8-31-2023

Pan-cancer proteogenomics connects oncogenic drivers to functional states

Yize Li
Washington University School of Medicine in St. Louis
Song Cao
Washington University School of Medicine in St. Louis
Matthew A Wyczalkowski
Washington University School of Medicine in St. Louis
Yizhe Song
Washington University School of Medicine in St. Louis
Erik P Storrs
Washington University School of Medicine in St. Louis

See next page for additional authors

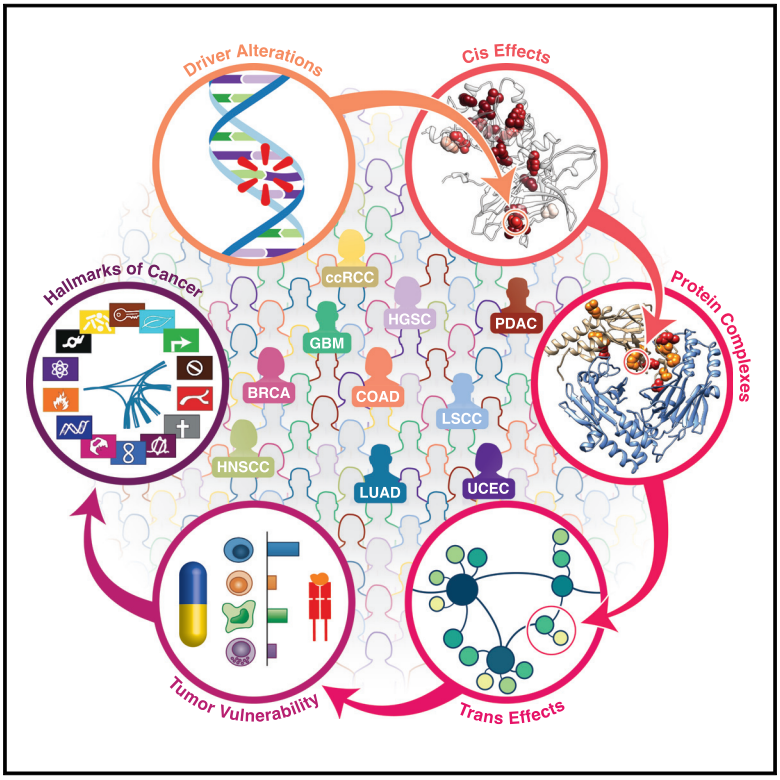
Recommended Citation

Li, Yize; Cao, Song; Wyczalkowski, Matthew A; Song, Yizhe; Storrs, Erik P; Wendl, Michael C; Liang, Wen-Wei; Terekhanova, Nadezhda V; Rodrigues, Fernanda Martins; Zhou, Daniel Cui; Wang, Liang-Bo; Baral, Jessika; Chheda, Milan G; Ding, Li; and et al., "Pan-cancer proteogenomics connects oncogenic drivers to functional states." Cell. 186, 18. 3921 - 3944.e25. (2023).
https://digitalcommons.wustl.edu/oa_4/3096

Cell

Pan-cancer proteogenomics connects oncogenic drivers to functional states

Graphical abstract



Highlights

- Multi-omic clusters reveal shared oncogenic driver pathways across ten cancer types
- Genetic changes correlate with altered, tumor-specific protein-protein interactions
- *cis/trans*-effects and kinase activities show driver heterogeneity and druggability
- Proteomic integration with genomic drivers resolves distinct cancer hallmark patterns

Authors

Yize Li, Eduard Porta-Pardo, Collin Tokheim, ..., Gad Getz, Li Ding, Clinical Proteomic Tumor Analysis Consortium

Correspondence

lewis_cantley@dfci.harvard.edu (L.C.C.), gadgetz@broadinstitute.org (G.G.), lding@wustl.edu (L.D.)

In brief

A multi-omics analysis-based resource across ten cancer types from more than 1,000 patients provides pan-cancer insights into shared oncogenic driver mechanisms and pathways.

Li et al., 2023, Cell *186*, 3921–3944
August 31, 2023
https://doi.org/10.1016/j.cell.2023.07.014

CellPress

Cell

CellPress
OPEN ACCESS

Article

Pan-cancer proteogenomics connects oncogenic drivers to functional states

Yize Li,[1,2,31] Eduard Porta-Pardo,[3,4,31] Collin Tokheim,[5,6,31] Matthew H. Bailey,[7,31] Tomer M. Yaron,[8,9,10,31] Vasileios Stathias,[11,12,32] Yifat Geffen,[13,14,32] Kathleen J. Imbach,[3,4,32] Song Cao,[1,2,32] Shankara Anand,[13] Yo Akiyama,[13] Wenke Liu,[15,16] Matthew A. Wyczalkowski,[1,2] Yizhe Song,[1,2] Erik P. Storrs,[1,2] Michael C. Wendl,[2,17,18] Wubing Zhang,[5] Mustafa Sibai,[3,4] Victoria Ruiz-Serra,[3,4] Wen-Wei Liang,[1,2] Nadezhda V. Terekhanova,[1,2] Fernanda Martins Rodrigues,[1,2] Karl R. Clauser,[13] David I. Heiman,[13] Qing Zhang,[13] Francois Aguet,[13] Anna P. Calinawan,[19] Saravana M. Dhanasekaran,[20] Chet Birger,[13] Shankha Satpathy,[13] Daniel Cui Zhou,[1,2] Liang-Bo Wang,[1,2] Jessika Baral,[1,2] Jared L. Johnson,[8,9] Emily M. Huntsman,[8,9] Pietro Pugliese,[21] Antonio Colaprico,[11,22] Antonio Iavarone,[11,23] Milan G. Chheda,[1,24,25]

(Author list continued on next page)

[1]Department of Medicine, Washington University in St. Louis, St. Louis, MO 63110, USA
[2]McDonnell Genome Institute, Washington University in St. Louis, St. Louis, MO 63108, USA
[3]Josep Carreras Leukaemia Research Institute (IJC), Badalona 08916, Spain
[4]Barcelona Supercomputing Center (BSC), Barcelona 08034, Spain
[5]Department of Data Science, Dana-Farber Cancer Institute, Boston, MA 02215, USA
[6]Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA 02115, USA
[7]Department of Biology and Simmons Center for Cancer Research, Brigham Young University, Provo, UT 84602, USA
[8]Meyer Cancer Center, Weill Cornell Medicine, New York, NY 10021, USA
[9]Department of Medicine, Weill Cornell Medicine, New York, NY 10021, USA
[10]Englander Institute for Precision Medicine, Institute for Computational Biomedicine, Weill Cornell Medicine, New York, NY 10021, USA
[11]Sylvester Comprehensive Cancer Center, University of Miami Miller School of Medicine, Miami, FL 33136, USA
[12]Department of Molecular and Cellular Pharmacology, University of Miami Miller School of Medicine, Miami, FL 33136, USA
[13]Broad Institute of Massachusetts Institute of Technology and Harvard, Cambridge, MA 02142, USA
[14]Cancer Center and Department of Pathology, Massachusetts General Hospital, Boston, MA 02115, USA
[15]Institute for Systems Genetics, NYU Grossman School of Medicine, New York, NY 10016, USA
[16]Department of Biochemistry and Molecular Pharmacology, NYU Grossman School of Medicine, New York, NY 10016, USA
[17]Department of Genetics, Washington University in St. Louis, St. Louis, MO 63130, USA
[18]Department of Mathematics, Washington University in St. Louis, St. Louis, MO 63130, USA
[19]Department of Genetic and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, NY 10029, USA
[20]Michigan Center for Translational Pathology, Department of Pathology, University of Michigan, Ann Arbor, MI 48109, USA
[21]Department of Science and Technology, University of Sannio, 82100 Benevento, Italy

(Affiliations continued on next page)

SUMMARY

Cancer driver events refer to key genetic aberrations that drive oncogenesis; however, their exact molecular mechanisms remain insufficiently understood. Here, our multi-omics pan-cancer analysis uncovers insights into the impacts of cancer drivers by identifying their significant *cis*-effects and distal *trans*-effects quantified at the RNA, protein, and phosphoprotein levels. Salient observations include the association of point mutations and copy-number alterations with the rewiring of protein interaction networks, and notably, most cancer genes converge toward similar molecular states denoted by sequence-based kinase activity profiles. A correlation between predicted neoantigen burden and measured T cell infiltration suggests potential vulnerabilities for immunotherapies. Patterns of cancer hallmarks vary by polygenic protein abundance ranging from uniform to heterogeneous. Overall, our work demonstrates the value of comprehensive proteogenomics in understanding the functional states of oncogenic drivers and their links to cancer development, surpassing the limitations of studying individual cancer types.

INTRODUCTION

Cancer is primarily initiated by genetic driver mutations in tumor suppressor genes (TSGs) and proto-oncogenes.[1] Multiple metrics are considered to define a gene as a cancer driver, including mutation recurrence in a gene,[2] functional protein domains,[3] individual amino acid hotspots for oncogenic mutations,[4–6] enrichment of damaging mutations,[7] or three-dimensional clustering of

Christopher J. Ricketts,[26] David Fenyö,[15,16] Samuel H. Payne,[27] Henry Rodriguez,[28] Ana I. Robles,[28] Michael A. Gillette,[13,30] Chandan Kumar-Sinha,[20] Alexander J. Lazar,[29] Lewis C. Cantley,[8,9,*] Gad Getz,[13,14,30,*] and Li Ding[1,2,17,24,33,*] Clinical Proteomic Tumor Analysis Consortium

[22]Department of Public Health Sciences, University of Miami Miller School of Medicine, Miami, FL 33136, USA
[23]Department of Neurological Surgery, Department of Biochemistry and Molecular Biology, University of Miami Miller School of Medicine, Miami, FL 33136, USA
[24]Siteman Cancer Center, Washington University in St. Louis, St. Louis, MO 63130, USA
[25]Department of Neurology, Washington University in St. Louis, St. Louis, MO 63130, USA
[26]Urologic Oncology Branch, Center for Cancer Research, National Cancer Institute, National Institutes of Health, Bethesda, MD 20892, USA
[27]Department of Biology, Brigham Young University, Provo, UT 84602, USA
[28]Office of Cancer Clinical Proteomics Research, National Cancer Institute, Rockville, MD 20850, USA
[29]Departments of Pathology & Genomic Medicine, The University of Texas MD Anderson Cancer Center, Houston, TX 77030, USA
[30]Harvard Medical School, Boston, MA 02115, USA
[31]These authors contributed equally
[32]These authors contributed equally
[33]Lead contact
*Correspondence: lewis_cantley@dfci.harvard.edu (L.C.C.), gadgetz@broadinstitute.org (G.G.), lding@wustl.edu (L.D.)
https://doi.org/10.1016/j.cell.2023.07.014

somatic mutations in the protein structure.[8–10] On applying these criteria to the large cohorts of cancer genomics data, the list of cancer genes and their predicted driver mutations has grown in recent years.[1,2,11] How these mutations mechanistically "drive" tumorigenesis remains incompletely understood.[12]

The Clinical Proteomics Tumor Analysis Consortium (CPTAC) has accelerated the understanding of basic molecular mechanisms in cancer by integrating data across the proteogenomic spectrum: whole-exome and whole-genome sequencing, DNA methylation, RNA-seq, and comprehensive proteomics and phosphoproteomics.[13–22] To date, extensive data have been generated for 1,000+ cases spanning ten cancer types: colorectal adenocarcinoma (COAD),[13] ovarian high-grade serous carcinoma (HGSC),[20] clear cell renal cell carcinoma (ccRCC),[14] head and neck squamous cell carcinoma (HNSCC),[18] lung squamous cell carcinoma (LSCC),[22] uterine corpus endometrial carcinoma (UCEC),[15] lung adenocarcinoma (LUAD),[17] pancreatic ductal adenocarcinoma (PDAC),[21] glioblastoma (GBM),[16] and breast carcinoma (BRCA).[19] Through interrogation of alterations in various omics layers, the effects of somatic driver mutations can be tracked to the unit of biological structure and function: the protein.

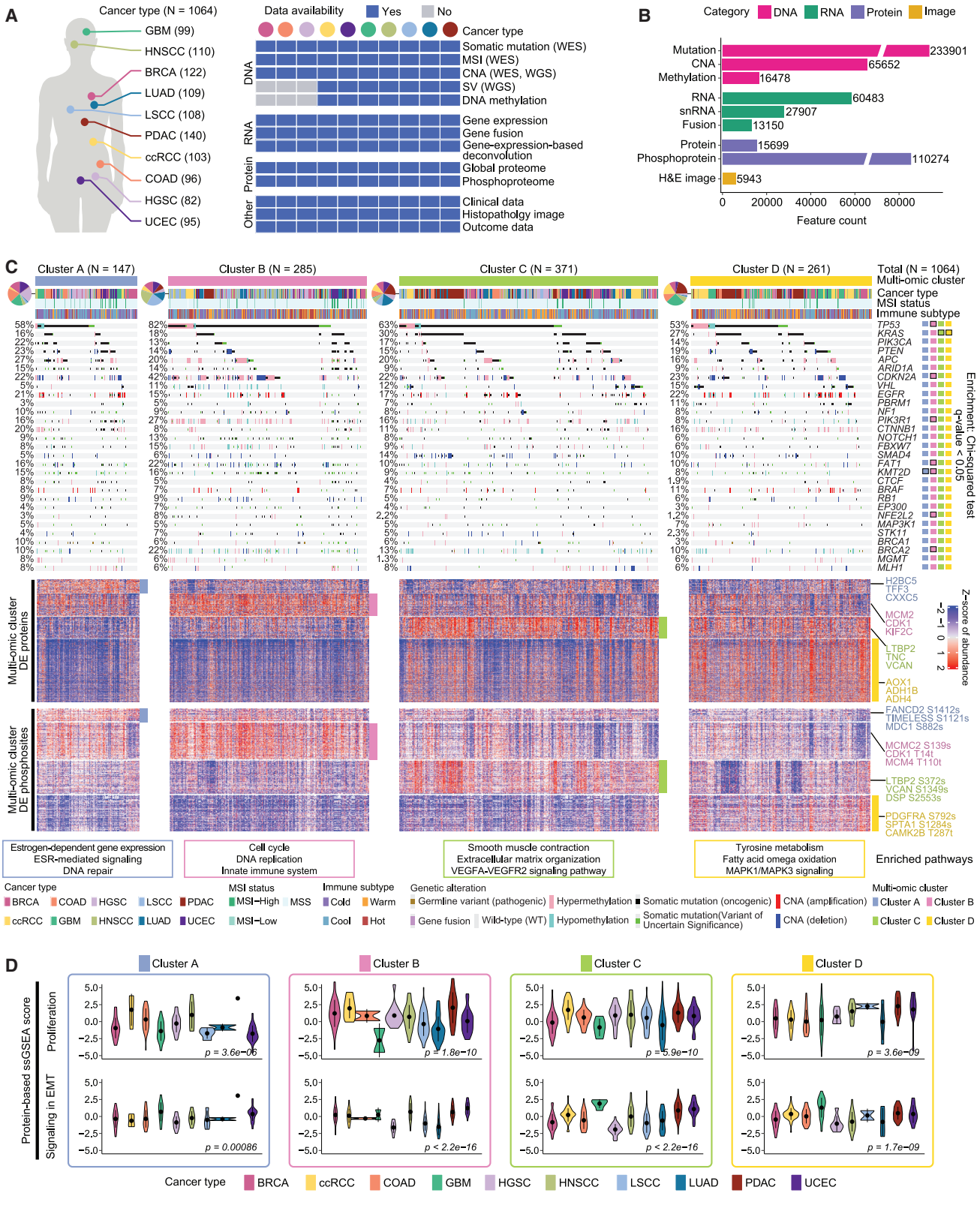
Pan-cancer studies focus on defining molecular characteristics across diverse cancers. Here, we extend our previous genomics-centered pan-cancer studies[1,23–26] by incorporating the proteomic layer to elucidate six critical aspects of cancer drivers: (1) pan-cancer genomic and epigenomic driver frequencies, exclusivities, and co-occurrences; (2) the impact of driver alterations on RNA, protein, and post-translational modifications (PTMs); (3) the effects of driver alterations on protein complexes; (4) essential protein- and phosphorylation-level changes in oncogenic pathways; (5) association of actionable driver alterations with the tumor microenvironment (TME); and (6) the combined effects of somatic drivers on protein abundance through the lens of cancer hallmarks. Our findings show the potential of integrative proteogenomic analysis in decoding oncogenic drivers and its potential clinical utility, particularly when a clear genomic target is absent.

RESULTS

A pan-cancer proteogenomic landscape of driver alterations and associated multi-omic clusters

Although large-scale DNA-sequencing studies have been vital in identifying cancer driver mutations, proteogenomics further enhances epigenomic, transcriptomic, and proteomic data to reveal functional consequences and therapeutic vulnerabilities. As part of CPTAC, we uniformly processed and analyzed proteogenomic data from 1,064 prospectively collected cases across ten cancer types, encompassing genetic alterations, DNA methylation, transcriptome, global proteomics, and phosphoproteomics (Figure 1A). Each case is accompanied by anonymized clinical data, histopathology, treatment outcome, and, in some cases, detailed molecular data from normal-adjacent tissues (NATs) (Table S1). A significant advance is the inclusion of mass-spectrometry-based proteomic and phosphoproteomic data, which exponentially expands the number of quantified features to 15,699 proteins and 110,274 phosphosites (Figure 1B), compared with 128 and 53, respectively, evaluated in The Cancer Genome Atlas (TCGA)'s RPPA-based assays.[27]

To reveal the landscape of tumor variability in the transcriptome, proteome, and phosphoproteome, we clustered tumor samples by applying Bayesian non-negative matrix factorization[28–30] followed by hierarchical clustering. We resolved four main multi-omic clusters (clusters A–D) (Figure 1C; STAR Methods). Cluster A lacks PDAC but is enriched with HGSC and COAD associated with a higher proportion of microsatellite instability (MSI-high) and immune-cold samples (p values = 0.02, 1e−5, respectively, chi-squared test; STAR Methods). Cluster B has more squamous cell carcinomas and higher immune infiltration than the other clusters (p value = 1e−5, chi-squared test). Clusters C and D are enriched in LUAD and GBM, respectively, and share similarities, such as a higher proportion of immune-warm samples (Figure 1C). Notably, some oncogenic alterations preferentially occurred in particular clusters (q value < 0.05, chi-squared test), such as *KRAS* mutations in clusters C and D, and *CDKN2A* and *TP53* mutations in cluster B (Figure 1C; Table S1). Moreover, we identified differentially



(legend on next page)

expressed proteins (DEPs) and phosphosites (DEPPs) (Figure 1C; Table S1). Functionally, cluster A-associated DEPs are enriched in estrogen-dependent gene expression and DEPPs in DNA repair. The DEPs and DEPPs in cluster B are enriched in DNA replication, innate immune system, and cell cycle, agreeing with the high number of *CDKN2A* alterations. Cluster C-associated DEPs and DEPPs relate to signaling in extracellular matrix (ECM) organization and VEGFA-VEGFR2. Finally, cluster D is enriched in tyrosine metabolism and MAPK1/MAPK3 signaling (Figure 1C).

Next, we quantified how representative pathways enriched in each multi-omic cluster varied by cancer type using protein-based single-sample gene set enrichment analysis (ssGSEA) (Figures 1D and S1A; Table S1). For example, ccRCC showed higher ssGSEA scores for the immune system pathway than GBM in each cluster (Figure S1A), consistent with a known higher immune infiltration and better response rate of ccRCC to immune checkpoint blockade.[31] To evaluate the clinical relevance, we conducted survival analyses among co-clustering tumors. Following multiple test corrections, we found that ccRCC samples in clusters C and D labeled as group A (i.e., enriched with PDAC samples) showed a superior prognosis to their counterparts (i.e., ccRCC samples in clusters A and B) (Figure S1B); this relationship was maintained after adjusting for age, sex, and tumor purity in a Cox proportional hazard model (p value = 0.008). By comparing the protein-based ssGSEA score of inflammatory response between groups A and B in ccRCC, we noticed that group A corresponded to lower immune infiltration levels (Figure S1C). A survival benefit was noted when we grouped the ccRCC samples based on their inflammatory response ssGSEA score (Figure S1D). GBM scored higher than other cancer types for the epithelial-mesenchymal transition (EMT) pathway, especially in cluster D (Figure 1D). The mesenchymal subtype of GBM expresses neural stem cell markers and is associated with an aggressive phenotype.[32–34] The subgroup of GBM with high EMT ssGSEA scores was linked to poor patient survival (p value = 0.0062; Figure S1E; STAR Methods). A holistic view of all cancer cases, based on their multi-omic signatures, enabled us to identify distinct sample clusters. However, how might the cancer drivers in these samples be related to the molecular pathway differences? We began by investigating the most apparent possible effect: the impact of an altered driver gene on its RNA and protein products.

Proteogenomic analyses reveal the heterogeneity of *cis*-effects of cancer mutations

Although genetic variants can broadly impact the proteome, the most proximal effect is at the transcriptional, translational, and post-translational product of the mutated gene itself.[35,36] To explore such *cis*-effects, we assessed how putative cancer driver gene mutations were associated with changes in their RNA, protein, or phosphoprotein levels with linear regression models (Figure 2A; Table S2; STAR Methods). We found 265 significant *cis*-events at the pan-cancer level in 59 cancer genes (q value < 0.1, Figure 2B; Table S2). As expected, TSGs such as *ARID1A*, *MSH6*, or *RB1* tended to be downregulated at the RNA, protein, or phosphoprotein level. Analyzed separately for each tumor type, we found 349 *cis*-effects in 59 cancer genes (q value < 0.1). Most overlapped with *cis*-effects from the pan-cancer analysis. Still, some were unique, including mutations in *CUL3* in LSCC, *NOTCH1* in HNSCC, and *KMT2C* in UCEC and BRCA, all associated with lower protein abundance (Table S2).

Extending this analysis to paired NATs provided additional insight into the dynamics of a subset of genes. Although the tumor-focused analysis showed that somatic mutations in *ARID1A* were associated with lower ARID1A protein abundance, NAT characterization revealed that tumors lacking *ARID1A* mutations, surprisingly, had higher ARID1A protein abundance than NATs (p value = 1e−16, Figures 2C and S2A). The reason for these elevated ARID1A protein levels in tumor samples without *ARID1A* mutations over matching NAT is unclear. Paired NATs also permitted further scrutiny of *STK11 cis*-effects. *STK11* mutations in LUAD correlated with diminished protein abundance. Wild-type (WT) *STK11* tumors exhibited higher STK11 protein abundance but had lower protein abundance than the matching NATs (p value = 1e−10, Figure 2D), suggesting that downregulation of STK11 is also crucial for lung tumors lacking *STK11* somatic mutations. Including NAT samples can contextualize how *cis*-effects alter protein levels relative to WT tumors and protein abundance in corresponding normal tissues.

Although certain types of variants are expected to have a specific *cis*-impact on protein levels (e.g., nonsense and frameshift indels), the consequences of other variants, like missense mutations, can vary widely. To assess whether missense mutations impact protein stability in cancer genes, we first normalized protein abundance by regressing out contributions from RNA expression (STAR Methods). After repeating the *cis*-effect analysis for putative oncogenic missense mutations, we found 11 pan-cancer and 15 cohort-specific significant events (q value < 0.05, Figure 2E; Table S2). As anticipated,[37] *TP53* missense mutations were associated with higher protein abundance, whereas frameshift indel and nonsense mutations were associated with lower protein abundance (Figure 2F). For all other TSGs with a *cis*-effect, oncogenic missense mutations were associated with lower protein abundance, as seen in *STK11*, *PBRM1*, and *PTEN* (Table S2). Consistent with an impact on protein stability, these missense mutations preferentially occurred at buried amino acids in the protein structure (p value =

Figure 1. The landscape of pan-cancer aberrations characterizes the four multi-omic clusters

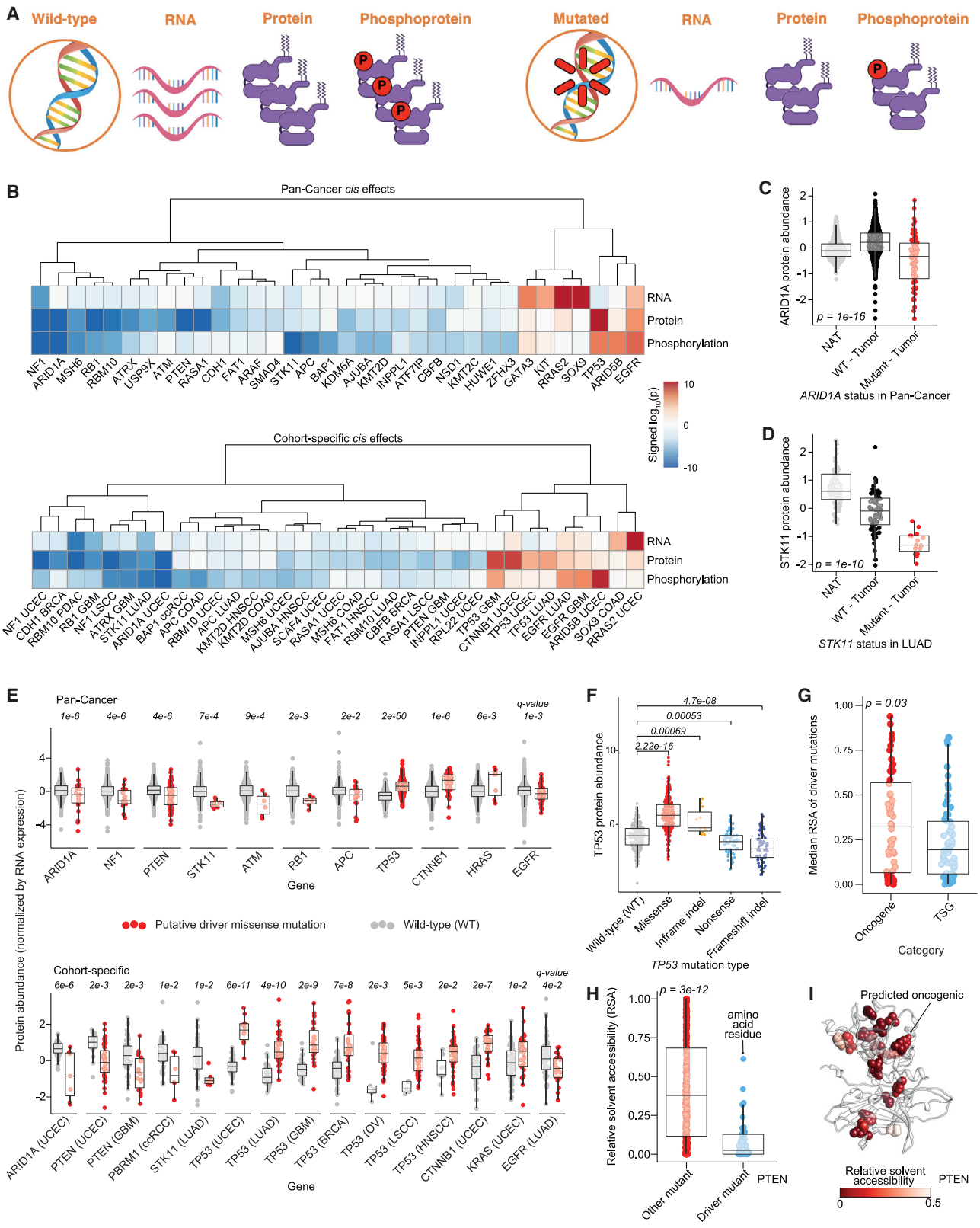
(A) The overview of cohort and data types.

(B) Quantified features in proteogenomics data types and the number of H&E images.

(C) We classified the tumor samples in this pan-cancer cohort into four multi-omic hierarchical clusters. They presented tumor variabilities in proteogenomics such as driver alterations (top), and the DEPs and DEPPs and enriched pathways associated with each cluster (middle and bottom).

(D) Violin plots showing the ssGSEA on the proteomic data in proliferation and EMT signaling pathways, grouped by multi-omic cluster. One-way ANOVA p values are shown, and data are represented as mean ± CI.

See also Figure S1 and Table S1.



(legend on next page)

0.03, Figure 2G; Table S2). For example, *PTEN* showed a substantial bias for putative driver missense mutations at buried amino acids (p value = 3e−12, Figure 2H; Table S2), especially for missense mutations computationally predicted as oncogenic (Figure 2I). Based on previous saturation mutagenesis screens,[38,39] these computationally predicted oncogenic mutations displayed low PTEN phosphatase activity like known oncogenic mutations from OncoKB (Figure S2B). However, unlike previously known oncogenic missense mutations, they exhibited lower protein abundance (Figure S2C). Uniquely enabled by a proteogenomics approach, these results highlight an underappreciated role of oncogenic missense mutations in lowering the protein abundance of some tumor suppressors (Figure S2D).

Inference of altered protein-protein interactions through protein co-variation analysis

Somatic mutations may modify protein-protein interactions (PPIs).[9,40] Although PPIs were not directly measured in CPTAC, known protein interaction pairs were noted to display high protein co-expression (Figure 3A).[41] We found higher correlation values for protein pairs that are indicated in larger numbers of PPI databases. For example, in the BRCA cohort, random protein pairs showed an average Pearson correlation coefficient of ~0, compared with 0.23 for PPIs represented in three or more databases (Figure 3B), such as those between MSH2 and MSH6 (Figure S3A). We used significant correlations between protein levels as indirect evidence of PPIs (STAR Methods).

First, the correlation between protein pairs from well-established PPIs was calculated involving at least one cancer gene across the CPTAC cohort (STAR Methods; Figure 3C). This revealed a set of conserved core PPIs, such as MSH2/MSH6, CTNNA1/CTNNB1, and RAD21/STAG1. Interestingly, most PPIs involving cancer genes only correlated in a subset of cohorts, suggesting tissue specificity. For instance, mTOR is known to interact with both RICTOR and RPTOR,[42] but we noted a correlation of mTOR protein abundance with RICTOR but not RPTOR in ccRCC, and the converse correlation with RPTOR but not RICTOR in GBM (Figure 3D). Similarly, CTNNB1 and CDH1 protein levels were seen to correlate in BRCA and UCEC but not in ccRCC and HGSC (Figure S3B). These results highlight the plasticity of PPIs across cancer types and the importance of measuring protein levels alongside RNA.

Next, we extended the concept of protein co-expression as an indicator of protein interactions to driver genes showing copy-number alterations (CNAs). We found 32 known interactors of driver genes significantly differentially expressed (q value < 0.1) relative to the CNA status of the driver gene (Figure 3E; Table S3), including deletion of *SMAD4* being associated with reduced SMAD2 protein levels and *CCNE1* amplifications associated with increased CDK2. We performed a mediation analysis to assess if the CNA-based *trans*-effect was indeed mediated through the protein abundance of the driver gene and not an unrelated gene within the CNA (Figures 3F and S3C; STAR Methods). We found that most PPI *trans*-effects (66%) were associated with protein abundance of the driver gene (q value < 0.1, natural effect model; Table S3). For example, copy-number amplifications in *RICTOR* are associated with increased protein abundance of MAPKAP1 (Figure 3G) and mTOR (Figure S3D) at the pan-cancer level, which could be deconvolved into a *cis*-effect on RICTOR protein levels (Figure 3H) and a *trans*-effect, irrespective of CNA status, between protein levels of RICTOR and MAPKAP1/mTOR (Figures 3I and S3E). Likewise, elevated CDK2 protein levels in *CCNE1*-amplified tumors could be deconvolved into a *cis*-effect and *trans*-effect (Figures S3F and S3G). At the phosphorylation-level, *CCNE1*-amplified tumors also showed elevated activity of the CDK protein sub-family containing CDK2 according to the Kinase Library (Figure S3H), which provides additional corroboration of our results.

Somatic mutations can change protein interactions by altering the interaction interface between proteins. Using an interaction term regression model (Figures S3I–S3K; STAR Methods), we tested whether putative oncogenic mutations in driver genes significantly changed the co-variation between protein interactors. In a pan-cancer analysis controlling for cancer type and tumor purity, we found 51 significant events (Figure 3J; Table S3)

Figure 2. *cis*-effects of SNVs in cancer genes across the CPTAC cohort

(A) An illustration visualizes the analysis structure for this figure. The effects of somatic mutations on RNA, protein, and phosphoproteomics were measured by comparing mutated and wild-type samples.

(B) Overview of the different *cis*-effects analyzed in this manuscript at the pan-cancer (top) and cohort-specific levels (bottom). Heatmaps showing the *cis*-effects for selected frequently mutated genes (x axis) at the RNA, protein, or phosphorylation levels (y axis) with effects measured at all three omics levels. Tiles are colored according to signed $\log_{10}$(p) of the *cis*-effect from red (positive) to blue (negative). Signed $\log_{10}$(p): multiplying the direction of the coefficient by the $\log_{10}$ of the adjusted p value (capped at ± 10).

(C) Pan-cancer comparison of the ARID1A protein levels (y axis) in tumors with *ARID1A* mutations (red) or without them (black) and NAT (gray), depending on the *ARID1A* mutation status (x axis).

(D) Comparison of the STK11 protein levels in LUAD patients (y axis) in tumors with *STK11* mutations (red) or without them (black) and NAT (gray) depending on the *STK11* mutation status (x axis). One-way ANOVA p values are shown, and data are represented as median and interquartile range in (Figures 2C) and (D).

(E) Analysis of oncogenic missense mutations associated with a *cis*-effect on protein abundance after adjusting for RNA expression level, either at the pan-cancer (top) or cohort-specific levels (bottom). Boxplots show tumor samples with a putative driver missense mutation (red) compared to wild-type tumor samples (gray). Wilcoxon rank-sum test q values are shown, and data are represented as median and interquartile range.

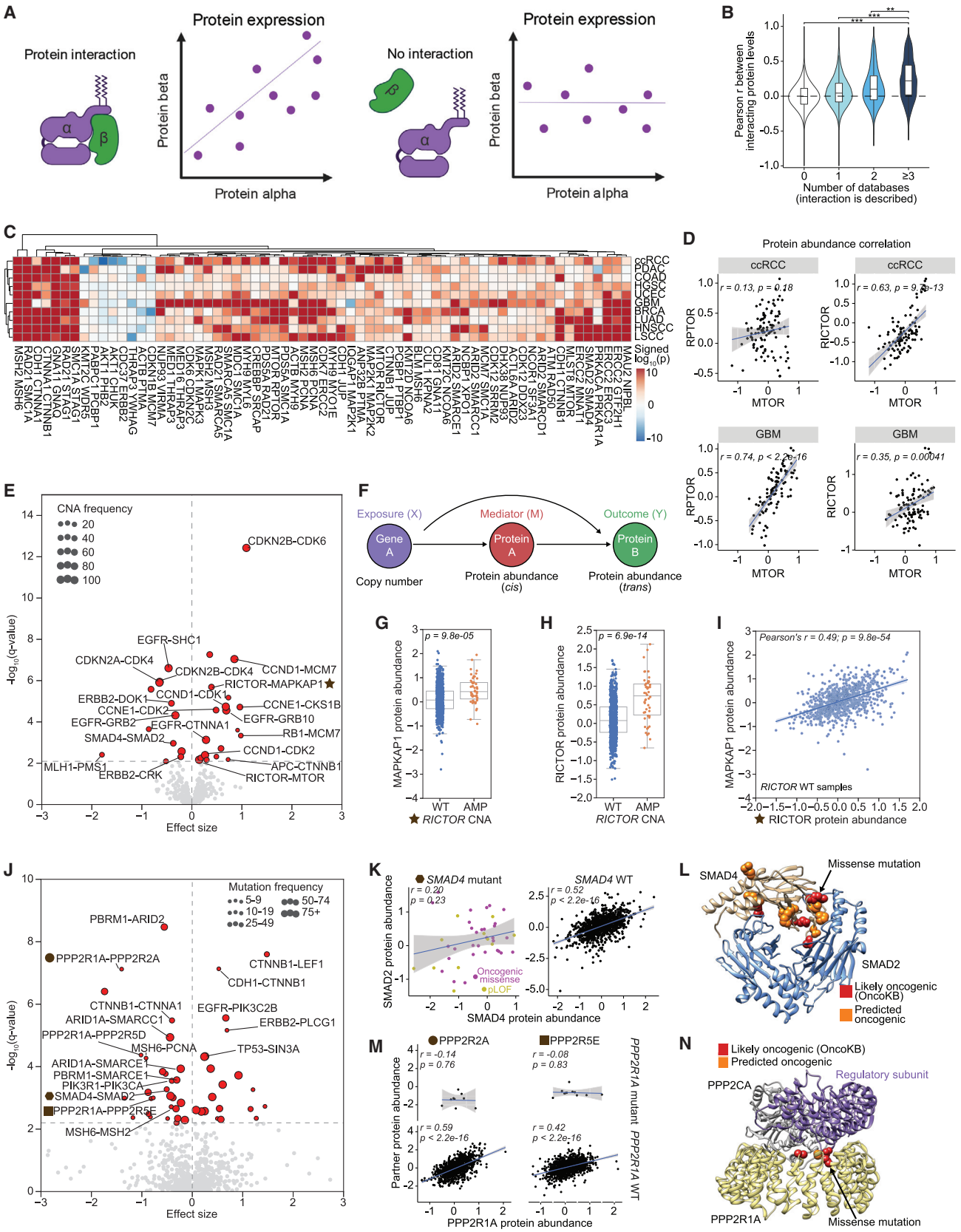
(F) Pan-cancer *cis*-effects of different types of *TP53* mutations (x axis) for TP53 protein abundance.

(G) Comparison of the relative solvent accessibility (RSA) for amino acid residues that have putative driver missense mutations for oncogenes (red) compared with TSGs (blue). RSA scores whether an amino acid is exposed at the surface of a protein (high score) or is buried (low score).

(H) Comparison of RSA for amino acids in PTEN with a putative driver missense mutation compared with all remaining amino acids. Wilcoxon rank-sum test p values are shown, and data are represented as median and interquartile range in (F)–(H).

(I) Protein structure showing the RSA for predicted oncogenic missense mutations in PTEN.

See also Figure S2 and Table S2.



(legend on next page)

affecting protein interactions. For example, oncogenic mutations in *CTNNB1* showed a substantially increased positive correlation with LEF1 (Figure S3L), a known effector of WNT/β-catenin pathway.[43] Conversely, oncogenic mutations in *PBRM1* abrogated the strong correlation of PBRM1 protein abundance with ARID2 (Figure S3M), which is consistent with the known dependence of PBRM1 protein stability on ARID2.[44] Similarly, several oncogenic point mutations in *SMAD4* were associated with a reduced correlation with SMAD2 (q value = 0.02; Figure 3K), suggesting a possible loss of protein interaction. This hypothesis is further supported by the localization of the missense mutations in SMAD4 at the protein interface with SMAD2 (Figure 3L). Lastly, oncogenic mutations in *PPP2R1A*, a scaffold protein for the phosphatase 2A holoenzyme, abrogated the correlation with a subset of regulatory subunits that control the specificity of the phosphatase, including PPP2R2A, PPP2R5E, and PPP2R5B (Figures 3M and S3N). The oncogenic missense mutations were similarly located at the interface of PPP2R1A and the regulatory subunits (Figure 3N). Notably, these observations were missed in RNA expression analysis, highlighting a unique advantage of proteomics (Figure S3O).

Finally, we investigated the potential influence of phosphorylation levels on protein interactions. At the pan-cancer level, 39 PPIs involving a driver were strongly influenced by phosphorylation levels (p value < 0.001), whereas 59 PPIs were affected at the cohort-specific level. Epidermal growth factor receptor (EGFR) phosphorylation at T693, particularly in ccRCC, appears to influence the correlation of EGFR and multiple protein partners (Figure S3P). 3D models of EGFR showed that T693 localizes at the intracellular interface between EGFR dimers, and phosphorylation at this site negatively impacts EGFR dimer formation (Figure S3Q).[45,46] Patients exhibiting high EGFR-T693 phosphorylation exhibit diminished correlations between EGFR protein and downstream EGFR-binding partners (Figures S3P and S3R). In conclusion, analyzing the impact of CNA, single-nucleotide variants (SNVs), and phosphorylation on protein co-variation is a potentially powerful method to understand interaction network rewiring in cancer.

trans-effects of somatic mutations in cancer genes across the CPTAC cohort

The impact of somatic mutations can extend beyond their own protein product's *cis*-effects or direct interactions. These more distant *trans*-effects illuminate the far-reaching molecular perturbations that can reverberate from driver gene alterations. The aggregations of all *trans*-effects of a cancer gene can be considered its "molecular fingerprint." We hypothesized that these molecular fingerprints could assess the similarity between different pairs of mutated cancer genes and infer functional relationships (Figure 4A). To identify these gene pairs, we calculated the correlation between each driver gene's proteomic and phosphoproteomic *trans*-effects signature with one another (Figure 4B; STAR Methods). Although most driver similarity scores indicated a slightly positive correlation (r > 0), a select few pairs exhibited more extreme positive and negative scores (Figure 4B). The two cancer genes with the most similar mutation *trans*-effects were *KEAP1* and *NFE2L2*. KEAP1 is known to bind and subsequently mark the transcription factor NFE2L2 for degradation.[47] The high global correlation between all *trans*-effects attributed to these proteins (r > 0.63, p value < 1e−15, Figure 4C) suggests that the overall cellular effect of mutations in either of these genes is the same: the enhanced protein stability of NFE2L2 and the consequent overexpression of its transcriptional targets. This is consistent with the fact that mutations in these genes are

Figure 3. Inference of altered protein-protein interactions through protein co-variation analysis

(A) Conceptual diagram showing that known protein-protein interactions tend to have higher protein co-expression than those not known to interact.

(B) Violin plot indicating the Pearson correlation between protein-protein abundance as the number of databases supporting a protein-protein interaction increases. Wilcoxon rank-sum test p values are shown. **p value < 0.01 and ***p value < 0.001. Data are represented as mean ± CI.

(C) Heatmap showing the correlation of protein abundance between cancer drivers and known protein interactors (x axis) across cancer types (y axis). Correlations are colored by their signed $\log_{10}$ p value, with red indicating positive and blue indicating negative correlations.

(D) Scatterplots of examples that indicate the correlation between protein abundance of cancer drivers and protein interactors varies by cancer type (left, mTOR and RPTOR; right, mTOR and RICTOR). Correlation test p values are shown.

(E) Volcano plot indicating proteins differentially expressed based on the presence of putatively oncogenic CNAs in cancer driver genes. Text labels indicate the cancer driver on the left and the protein interactor on the right. The horizontal dashed line indicates the threshold for statistical significance (q value < 0.1).

(F) Diagram indicating that a mediation analysis tries to identify whether the *trans*-effect (outcome, Y) of a CNA (exposure, X) is mediated through altered protein abundance of the putative cancer driver gene (mediator, M).

(G) Boxplot indicating the upregulation of MAPKAP1 protein abundance in *RICTOR*-amplified tumors (*trans*-effect).

(H) Boxplot indicating the upregulation of RICTOR protein abundance in *RICTOR*-amplified tumors (*cis*-effect). The Wald test assesses statistical significance and represents data as median and interquartile range in (G) and (H).

(I) Scatterplot showing the correlation of RICTOR and MAPKAP1 protein levels in *RICTOR* wild-type tumor samples. The Wald test assesses statistical significance.

(J) Volcano plot indicating proteins with either increased or decreased association with a cancer driver when the tumor contains a putative oncogenic mutation.

(K) Scatterplot showing the correlation between SMAD2 and SMAD4 protein levels, stratified by whether the tumor has an oncogenic mutation in *SMAD4*. Correlation test p values are shown.

(L) Protein structure (PDB: 1U7V) of SMAD4 (tan) in complex with SMAD2 (blue). Mutated amino acid residues that are putatively oncogenic are shown in a spherical representation.

(M) Scatterplot showing the correlation between phosphatase 2A regulatory subunits (PPP2R2A and PPP2R5E) and PPP2R1A protein levels, stratified by whether the tumor has an oncogenic mutation in *PPp2R1A*. Correlation test p values are shown. r represents the Pearson correlation coefficient in (D), (I), (K), and (M).

(N) Protein structure of the phosphatase 2A holoenzyme (PDB: 2NPP), consisting of PPP2R1A (yellow), PPP2CA (gray), and a regulatory subunit (purple). Mutated amino acid residues that are putatively oncogenic are shown in a spherical representation.

See also Figure S3 and Table S3.

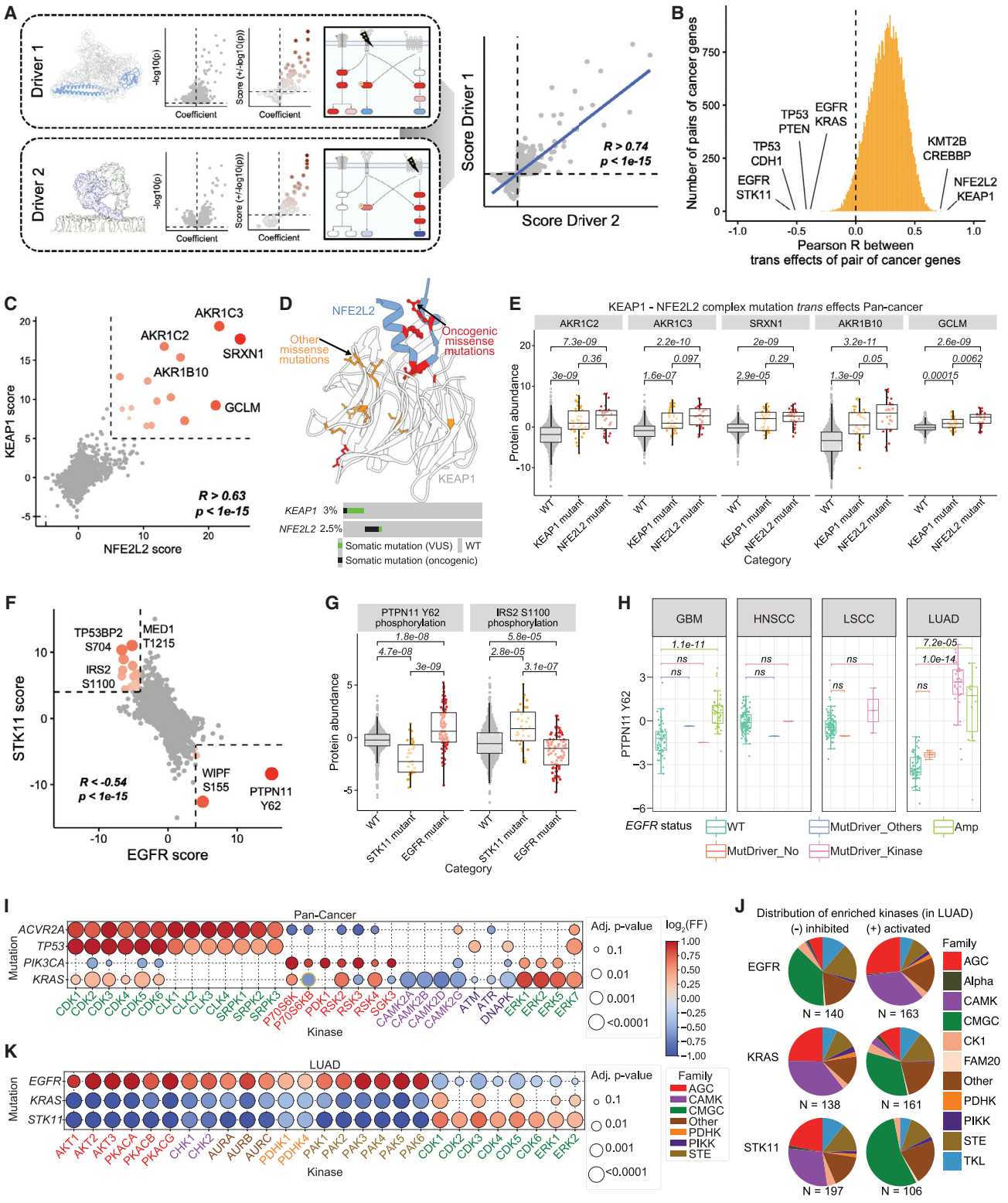


Figure 4. *Trans*-effects of somatic mutations in cancer genes across the CPTAC cohort

(A) Overview of how the similarity between *trans*-effects of driver genes was calculated based on proteomics and phosphoproteomics data from CPTAC.

(B) Histogram showing the distribution of Pearson r values of *trans*-scores (i.e., signed $\log_{10}(p)$) between all possible pairs of cancer genes in CPTAC (x axis).

(C) Scatterplot showing the *trans*-effects scores (i.e., signed $\log_{10}(p)$) for KEAP1 (y axis) and NFE2L2 (x axis).

(legend continued on next page)

mutually exclusive (Figure 4D). The strongest *trans*-effects of both genes are the upregulation of NFE2L2 targets such as AKR1C2 and AKR1C3 (Figure 4E).

Although most cancer genes showed positive correlations, suggesting similar *trans*-effects, we also observed negative correlations such as those between TP53/PTEN, TP53/CDH1, EGFR/KRAS, and EGFR/STK11, with the last example having the strongest negative correlation ($r < -0.54$, p value $< 1e{-}15$, Figure 4F). The EGFR/STK11 pair has several *trans*-effects that are opposite depending on whether the cancer cell has an *EGFR* or *STK11* mutation, including several phosphorylation sites such as TP53BP2 S704 and IRS2 S1100 (both increased in *STK11*-mutated and decreased in *EGFR*-mutated samples) or WIPF S155 and PTPN11 Y62 (increased in *EGFR*-mutated instances and vice versa, Figure 4G). Furthermore, most of these events are statistically significant in both directions compared with WT samples for both *EGFR* and *STK11* (Figure 4G). Notably, most of these effects were independent of the specific location of the mutation within *EGFR* (Figure 4H). These results suggest that *EGFR* and *STK11* mutations could push normal cells into opposite oncogenic states.

Next, we investigated the *trans*-effects of mutations on the phosphorylation state of the cell at the pan-cancer level. We compared the predicted kinase activities between mutated and WT samples across all cancer types (Figures 4I and S4B; Table S4). For example, mutations in both *ACVR2A* and *TP53* exhibited activation of cell cycle and splicing kinases; when mutated, the well-known oncogene *PIK3CA* showed strong activations of the PI3K/AKT pathway members S6Ks, RSKs, PDK1, and SGK3; mutations in *KRAS* led to a significant activation of its downstream extracellular signal-regulated kinases 1/2/5/7 (ERK1/2/5/7) (Figure 4I). Interestingly, we found that the CAMK2 kinases were inhibited in *KRAS*-mutated samples, concordant with previous studies.[48] In a pan-cancer analysis, *EGFR*-mutated samples showed an inverse kinase activity pattern to those of *STK11*- and *KRAS*-mutated samples (Figure S4B).

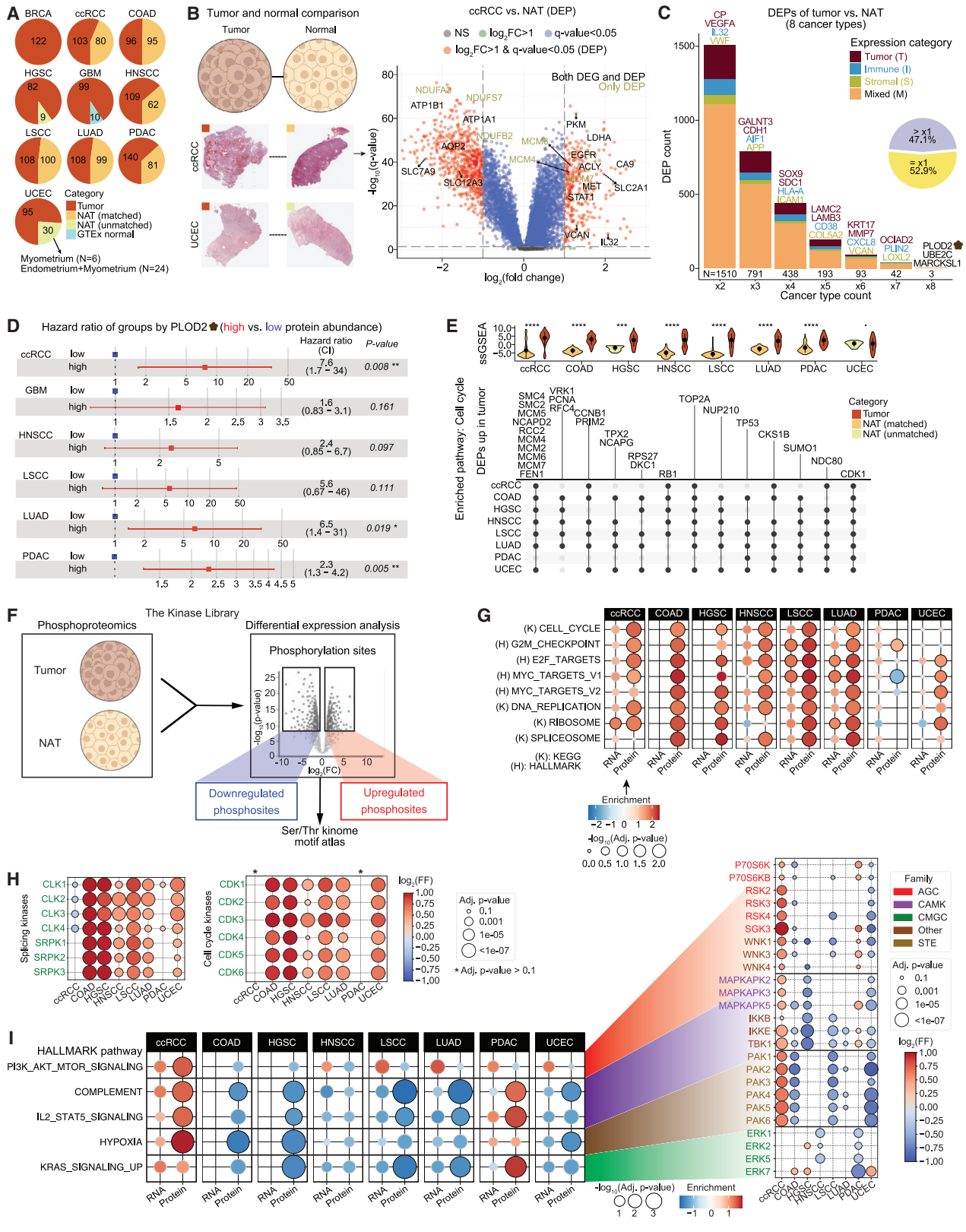
Furthermore, we used the Kinase Library to determine the activity state of different kinases in *EGFR*-mutated and *STK11*-mutated LUAD samples. We found that most activated kinases are from the CAMK and AGC families, whereas most inhibited kinases are from the CMGC family in *EGFR*-mutated samples (Figure 4J). In agreement with the observed negative correlation for the overall *trans*-effects, an opposite pattern was observed for both *STK11*-mutated and *KRAS*-mutated samples, where the CMGC family dominated the activated kinases, and the inhibited kinases were primarily members of the CAMK and AGC families (Figure 4J). Multiple kinases from different families showed opposite activities between *EGFR*-mutated samples and *STK11*-mutated/*KRAS*-mutated samples (Figures 4K and S4C; Table S4). These data suggest that *EGFR* mutation and *STK11/KRAS* mutations lead to distinct phosphorylation readouts in cancer cells. Our kinase enrichment analysis using substrate motifs around phosphosites can capture underlying mechanisms of driver mutations in cancer, informing treatment strategies beyond genetic testing. This has implications for basket clinical trial design and personalized oncology drug repurposing.

Comparative analysis between tumor and normal-adjacent tissue identifies key protein changes for oncogenic pathways

In contrast to other large-scale tumor characterization studies, eight of the ten CPTAC cohorts included matched (similar cell lineage) or unmatched (non-similar cell lineage) NATs (n = 556), and the GBM study contained ten Genotype-Tissue Expression (GTEx) project normal samples (Figure 5A). This allowed us to investigate divergent expression patterns between tumors and paired NATs. We found that tumors present distinct proteomic and phosphoproteomic signatures compared with their homologous NATs, such as in ccRCC (Figure 5B; Table S5), and identified 6,517 DEPs elevated in tumors and 7,030 DEPs elevated in NATs within the pan-cancer cohort. Of these 6,517 DEPs, 3,070 were differentially expressed in ≥ 2 cancer types, and the rest in a single cancer type. Among common DEPs, we found PLOD2, UBE2C, and MARCKSL1 in all cancer types (Figure 5C; Table S5). Notably, we observed that high PLOD2 protein abundance was significantly associated with worse overall survival in ccRCC, LUAD, and PDAC, and a similar trend in patients with GBM, HNSCC, and LSCC (Figure 5D). This suggests the potential value of PLOD2 as a pan-cancer prognostic biomarker.

To further investigate which cell types in the TME might express these proteins, we used single-nuclei RNA-seq (snRNA-seq) data and, based on correlation analysis with xCell

(D) Three-dimensional structure of the protein complex between KEAP1 and NFE2L2 (top, PDB 3WN7) and an oncoplot showing the CPTAC samples with alterations in these two genes.
(E) Boxplots showing the detailed protein levels (y axis) for each CPTAC sample of four proteins regulated by NFE2L2 (AKR1C2, AKR1C3, SRXN1, and AKR1B10) stratified by mutation status of *KEAP1* and *NFE2L2* of the tumor (x axis).
(F) Scatterplot showing the *trans*-effects of *STK11* mutations (y axis) or *EGFR* mutations (x axis) on all measured proteins and phosphoproteins. Correlation test p values are shown, and r represents the Pearson correlation coefficient in (A), (C), and (F).
(G) Boxplots showing the detailed protein levels (y axis) for each CPTAC sample of two phosphorylation events (PTPN11 pY62 and IRS2 pS1100) based on the mutation status of *EGFR* and *STK11* of the tumor (x axis).
(H) PTPN11 Y62 phosphorylation levels (y axis) in different CPTAC cohorts depending on the *EGFR* mutation status (x axis). Wilcoxon rank-sum test p values are shown, and data are represented as median and interquartile range in (E), (G), and (H).
(I) Pan-cancer kinase activity in tumors with specific mutations based on the Kinase Library. For each kinase and condition, the bubble color ($\log_2$FF) indicates the $\log_2$-ratio of the prediction frequency between the mutated tumors compared with WT tumors, and bubble size denotes the statistical significance by Fisher's exact test.
(J) Protein family distribution of kinase activity inferred from the Kinase Library in LUAD.
(K) The Kinase Library analysis of kinase activity in *EGFR*-mutated, *KRAS*-mutated, and *STK11*-mutated LUAD tumors.
See also Figure S4 and Table S4.



(legend on next page)

and ESTIMATE signatures, classified the proteins into four categories: tumor (T), immune (I), stromal (S), and mixed (M) (Figures 5C and S5A; Table S5; STAR Methods). This allowed us to evaluate the dysregulated signatures in TME at a higher resolution. Pathway analysis of DEPs enriched in individual categories revealed that pathways of keratinization, EGFR, and Warburg effect in the T subset; immune system and nonsense-mediated mRNA decay (NMD) in the I subset; and ECM organization and collagen formation in the S subset, reflecting the distinct characteristics of TME components (Table S5). Furthermore, we observed that MSI-high tumors showed significantly higher NMD pathway scores than microsatellite-stable (MSS) tumors (Figures S5B and S5C). As reported, MSI COAD tumors expressed high levels of critical activators of the NMD system and that inhibition of NMD *in vivo* using amlexanox reduced MSI tumor growth.[49] These results suggest that inhibition of the oncogenic activity of NMD may provide an additional therapeutic avenue for the personalized treatment of MSI tumors. Besides, the protein-based cell-cycle ssGSEA score was significantly elevated in tumors relative to NATs in seven of eight cancer types (Figure 5E). The difference was not significant in UCEC, possibly due to NATs of UCEC consisting mainly of the myometrium with various levels of admixed endometrium or because these non-tumor samples were derived from women at different stages of their menstruation cycle (Figure 5A). Among cell-cycle pathway proteins, we found 27 upregulated DEPs in multiple cancer types (Figure 5E). Similarly, we detected a consistently enriched pathway of transcriptional regulation by TP53 in tumors compared with NATs across cancer types (Figure S5D). As for the DEPs evaluated in specific cancer types, they contributed to pathways such as the metabolism of lipids in UCEC and keratinization in HNSCC (Figure S5E).

Finally, we used the Kinase Library to infer kinase activity patterns in different cancers from our phosphoproteomics data (Figure 5F; STAR Methods). We first applied protein- and RNA-based pathway enrichment analysis to characterize shared or distinct pathways across tissues. Interestingly, cell cycle and other replication-related pathways were enriched in most tumor types, predominantly only at the protein level (Figure 5G; Table S5). The Kinase Library showed broad activation of multiple CDKs in most tumor types, except ccRCC and PDAC (Figure 5H). Splicing pathways were also predominantly observed at the protein level in various tissues (Figure 5G), consistent with activated splicing kinases in the corresponding tissues (Figure 5H). Furthermore, we found agreement between kinase activity in distinct tissues and their related enriched pathways at the protein level, such as the PI3K-AKT-mTOR pathway was enriched only in ccRCC (Figure 5I; Table S5). Moreover, we observed a potential feedback mechanism for KRAS signaling in PDAC that was only evident at the phosphorylation level, where upregulation of KRAS signaling was associated with reduced activity of the ERK family protein kinases. This insight may explain the reduced clinical efficacy of therapeutic agents in *KRAS*-mutated pancreatic tumors[50] and further emphasize the importance of proteomic and phosphoproteomic characterization of cancer (Figure S5F).

Impact of somatic mutations on immunogenic neoantigens and druggable kinases

Given the relevance of neoantigens to immunotherapy,[51] we systematically predicted neoantigens that bind to the patient-specific human leukocyte antigen (HLA) class I alleles (STAR Methods). We found the expected positive correlation of neoantigen burden with tumor mutation burden[52] and higher neoantigen burden noted in tumor samples with mutational signatures of smoking and MSI (Figure 6A). The predicted neoantigen burden normalized by mutational burden was higher for MSI-high samples, likely due to a higher frequency of insertions and deletions resulting in frameshifts (Figure S6A). Neoantigen burden was positively correlated with inferred T cell infiltration (p value = 4e−4, cytotoxic T lymphocyte (CTL) score; p value = 0.002, CIBERSORTx CD8 T cells; Wald test) at the pan-cancer level (Figure S6B). However, only cancer types that tended to have a high baseline tumor mutation burden (COAD, UCEC, LUAD, and BRCA) were nominally significant by themselves, suggesting a minimal neoantigen burden threshold for immunogenicity.

To investigate the relationship between predicted neoantigen burden and T cell infiltration in the context of specific driver mutations, we employed an interaction term regression model (STAR Methods). Oncogenic alterations in seven driver genes were associated with a change in the correlation between neoantigen burden and inferred T cell infiltration, including KRAS (q value < 0.1; Figure 6B; Table S6). Investigating the potential correlation between neoantigen expression level and immunogenicity, we observed that a subset of genes that harbored a high number of predicted neoantigens (≥20) in MSI-high tumor

Figure 5. Analysis of normal-adjacent tissue identifies key protein changes for oncogenic pathways

(A) The distribution of tumors and normal samples across cancer types.

(B) The schematic of the tumor and normal comparison is represented by selected H&E staining images of the tumor and normal pairs from ccRCC and UCEC cohorts. The volcano plot shows such differentially expressed genes (DEGs) and differentially expressed proteins (DEPs) in ccRCC as an example.

(C) The distribution of DEP identified by tumor and NAT comparison. The pie chart indicates the distribution of cancer-type-specific DEPs and the rest.

(D) The plot shows that groups with higher PLOD2 protein abundance present increased risk in survival reflected by significantly higher hazard ratios in multiple cancer types. Cox proportional hazards model p values are shown. Whiskers represent ± 95% confidence intervals of the hazard ratios.

(E) Cell-cycle protein-based ssGSEA scores are significantly higher expressed in tumors than in NATs in seven of eight cancer types. Wilcoxon rank-sum test p values are shown. •p value > 0.05; ***p value < 0.001; and ****p value < 0.0001. Data are represented as mean ± CI.

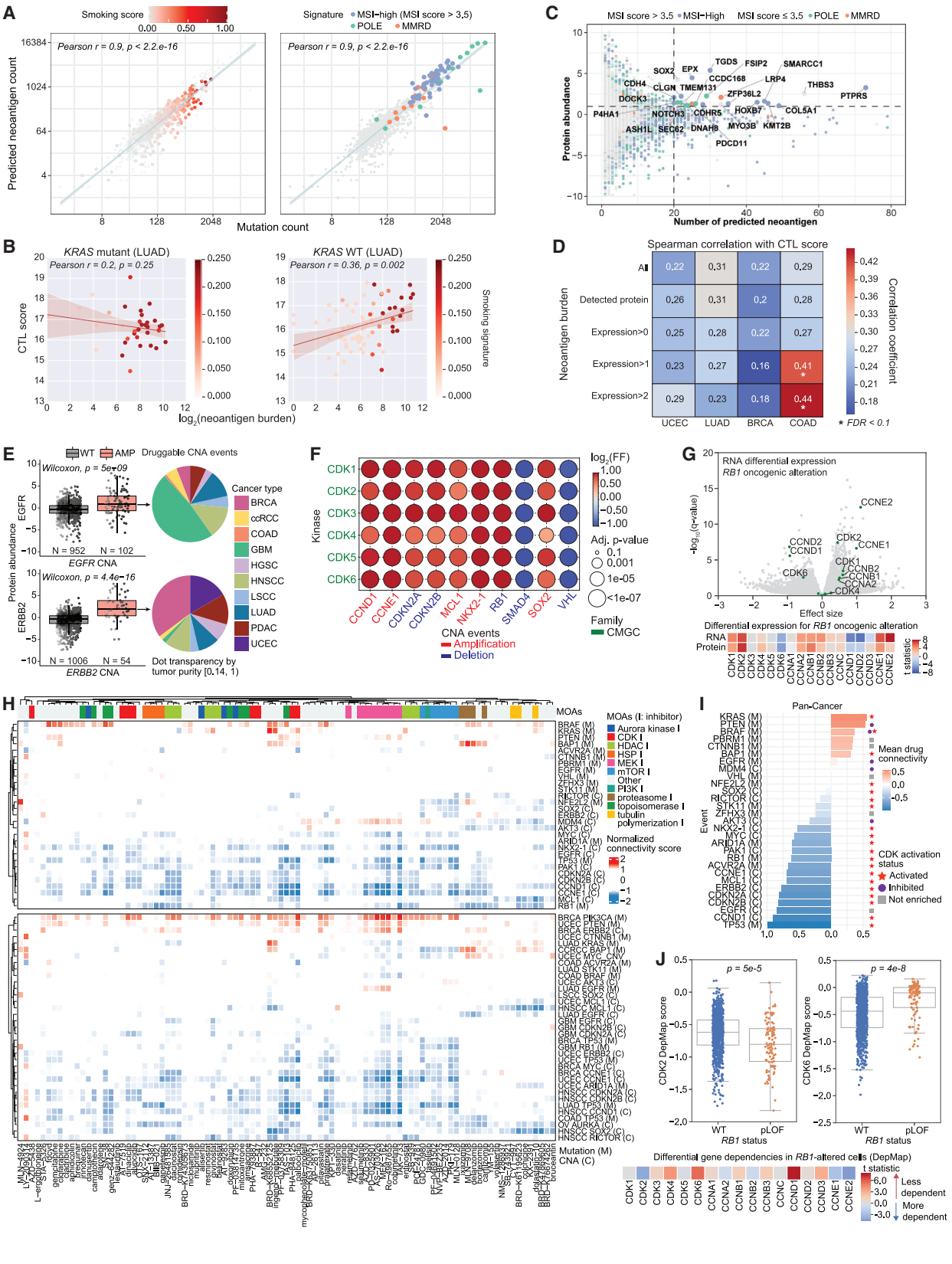
(F) Workflow of phosphoproteomics analysis based on the Kinase Library to reveal distinct activity patterns of protein kinases in tumor and normal comparison across cancer types.

(G) Pathway analysis (GSEA) at RNA and protein levels.

(H) The Kinase Library analysis for activated kinases in tumors compared with NAT (homologous or non-homologous) across multiple tissues.

(I) Pathway analysis (GSEA) and kinase activity analysis of tissue-specific trends.

See also Figure S5 and Table S5.



(legend on next page)

samples were also highly expressed at the protein level (protein abundance > 1) (Figure 6C). The neoantigen burden in highly expressed proteins vs. all proteins improved the correlation with inferred T cell infiltration (q value = 0.03, likelihood ratio test, Figure 6D) in COAD, where MSI is common. Notably, the neoantigen burden for highly expressed transcripts vs. all transcripts did not improve correlations (Figure S6B), highlighting an advantage of proteomics.

Although copy-number amplifications for oncogenes were associated with higher protein abundance than WT tumors, as was expected (Figure S6C), they also showed higher variance in protein abundance (p value = 0.006, Mann-Whitney U test, Figure S6D; Table S6). These findings extended to druggable targets (Figure 6E). *EGFR*- and *ERBB2*-amplified tumors had higher protein abundance than their WT counterparts (p value = 4e−16 and 5e−9, respectively). However, there is still considerable overlap in the distribution of protein abundance in *cis* (Figure 6E), as the average precision for distinguishing *ERBB2* and *EGFR*-amplified from WT tumors is only 0.53 and 0.37 (of 1.0), respectively. Protein abundance might provide additional information beyond the genomic event alone for therapeutic decisions. Because a limited number of oncogenes are directly targetable, an analysis of the *trans*-effects of a CNA might further expand potentially druggable targets. Using the Kinase Library, we analyzed CNA events that might activate druggable kinases (Figure S6E). We found that multiple CNA events led to the activation of CDKs (Figure 6F; Table S6), including the deletion of *CDKN2A*, which has previous experimental and clinical support as a potential biomarker for CDK4/6i treatment.[53–55] Beyond *CDKN2A* deletion, we found additional CNA events within and outside the core cell-cycle pathway associated with CDK activation (Figures 6F and S6E). Tumors with oncogenic alterations for *RB1* were associated with marked upregulation of CDK2 at both the RNA and protein levels (Figure 6G; Table S6), possibly related to the loss of *RB1* and its known role in transcriptional repression (Figures S6F and S6G).[56] In contrast, *RB1*-altered tumors also showed substantial downregulation of CDK6 at both the RNA and protein levels (Figure 6G). Thus, the signal of CDK activation from the Kinase Library may reflect the activation of distinct CDKs depending on the driver alteration.

To distinguish scenarios where CDK activation represents a cancer dependency, we expanded our analysis of potential therapeutic vulnerabilities by analyzing cell-line drug-response data concerning the proteomic signatures of driver alterations, including somatic mutations and CNAs. We calculated normalized drug connectivity scores between proteomic signatures of 64 driver alterations and drug-response signatures (CMAP Library of Integrated Network-Based Cellular Signatures [LINCS] database[57]) for compounds that are either FDA-approved or under active clinical investigation (Figure 6H; Table S6; STAR Methods). Interestingly, in agreement with the Kinase Library results (Figures 6F, S5C, and S6F), we found pan-cancer proteomic signatures related to genes within and outside the cell-cycle pathway to be associated with drug connectivity to multiple CDK inhibitors (CDKi) (Figure 6H, top). These findings could also be seen for tissue-specific signatures of driver alterations (Figure 6H, bottom).

In some cases, such as for *RB1* deletions, strong CDK activation was seen in the Kinase Library but a relatively lower association with CDKi drug-response profiles (Figure 6I). This may be related to a known role for *RB1* mutations in primary and acquired resistance to CDK4/6i, such as palbociclib.[58,59] Consistent with this, examining CRISPR gene KO screens (DepMap)[60] showed *RB1* altered cancer cell lines were less dependent on CDK4/6 but more dependent on CDK2 (q value < 0.1, Wald test; Figures 6J, S6H, and S6I; Table S6). Although not included in CMAP, recently developed small molecules selective for CDK2[61] could be more effective in *RB1*-altered tumors. Cumulatively, most driver alterations that showed CDK activation in the Kinase Library exhibited high

Figure 6. Immunogenicity and druggability analysis of somatic alterations

(A) Pearson correlation of the number of mutations in a tumor with neoantigen burden, stratified by mutational signatures (Smoking signature weights) or DNA repair deficiency (POLE, POLE mutant; MMRD, mismatch repair deficient).

(B) Pearson correlation of neoantigen burden and cytotoxic T lymphocyte (CTL) score stratified by driver mutation status of *KRAS* in LUAD. Wald test p value is shown.

(C) Scatterplot indicating genes with more predicted neoantigens simultaneously present in highly expressed proteins. Dashed lines indicate > 1 protein abundance (y axis) and > 20 neoantigens (x axis).

(D) Heatmap showing Spearman's correlation of CTL score with neoantigen burden using different subsets of proteins based on protein abundance.

(E) Boxplot indicating the *cis*-effect of CNAs on protein levels for two druggable targets. Wilcoxon rank-sum test p values are shown, and data are represented as median and interquartile range.

(F) Kinase activity for different CNAs at the pan-cancer level. For each kinase and condition, the color of the bubble (log_2FF) indicates the log_2-ratio of the prediction frequency for that kinase between the tumors with the CNA compared to the WT tumors, and the size of the bubble denotes the statistical significance by Fisher's exact test.

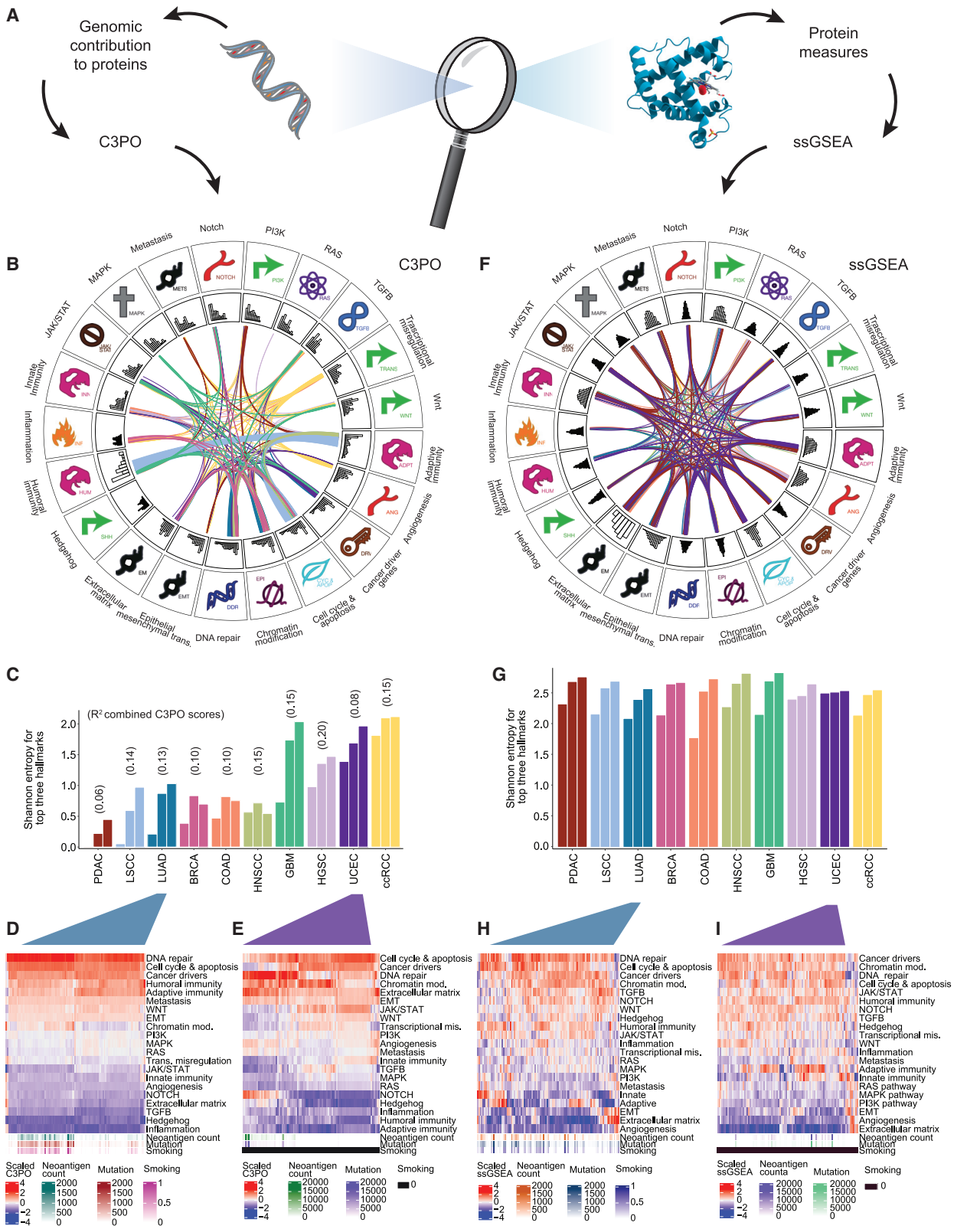
(G) Top, volcano plot showing the DEGs of tumors containing oncogenic *RB1*-alterations. Green dots indicate cell-cycle-related genes (CDK1-6). Bottom, heatmap showing the differential expression at the RNA and protein level for cell-cycle related genes (Wald test). Positive t statistic (red) indicates upregulation; negative t statistic (blue) indicates downregulation.

(H) Heatmap showing the normalized drug connectivity scores between protein abundance signatures (derived from driver alterations) and the drug-response signatures from the CMAP LINCS 2020 dataset. Negative connectivity scores (blue) indicate the predicted sensitivity of a signature to a particular drug.

(I) Average drug connectivity between seven CDK inhibitors used in the L1000 assay and pan-cancer (top) and cancer-specific (bottom) driver alteration signatures. CDK activation status was inferred using the Kinase Library.

(J) Top, boxplots showing cancer cell-line dependencies for two genes (CDK2 and CDK6) from CRISPR KO screens (DepMap) depending on the presence of pLOF mutation in the *RB1* gene. Wilcoxon rank-sum test p values are shown, and data are represented as median and interquartile range. Bottom, heatmap showing the differential gene dependencies (t statistic, Wald test) for cell-cycle-related genes (CDK1-6).

See also Figure S6 and Table S6.



(legend on next page)

(negative) drug connectivity to CDKi by using independent proteomic signatures (Figure 6I). These results suggest the potential therapeutic benefits of CDKi for treating tumors harboring genomic alterations (*MCL1* or *ERBB2* amplification) that are not necessarily directly linked to the cell-cycle pathway.

Integrative multi-genomic scoring provides evidence on how somatic alterations alter cancer hallmarks

Individual gene alteration can impact the proteomic landscape in *cis*, *trans*, and mediator settings. However, we wished to explore the global effect of all somatic mutations on the overall protein variability within cancer types. To address this question, we built a combined polygenetic protein abundance prediction algorithm for oncology, called C3PO, as a theoretical exercise. This tool applies polygenic mathematics to describe protein variability instead of disease status, like polygenic risk scores (PRS).[62–64] To calculate a polygenic predictor, one measures changes in protein abundance in the altered or unaltered state (Figure S7A). The effect sizes can be summed across multiple alterations to produce a polygenic risk score based on which alterations are present. Typical applications of PRS generate effect sizes by comparing mutated samples to non-mutated samples for a phenotype of interest. However, because of the large number of infrequently mutated driver genes in cancer (Figure S1E), we aggregated mutations into NeSTed protein systems[65] (Table S7; STAR Methods). Doing so provided broader coverage of the possible mutational impacts on protein. Results from C3PO suggest that somatic mutations and copy-number events alone account for, on average, 7.2%–27.5% (range 0%– 91.9%) of the protein abundance variability across these ten cancer types (Figures S7B and S7C). When applied to an orthogonal dataset, C3PO could only collectively account for 0.7%–2.6% of global protein variability in HGSC (Figures S7D and S7E; 1.5% CNA amplification, 1.5% CNA deletion, and 0.8% somatic mutations, range 0%–19%),[20] BRCA (2.0% CNA amplification, 2.0% CNA deletion, and 2.3% somatic mutations, range 0%–31.2%),[19] and COAD (2.2% CNA amplification, 2.0% CNA deletion, and 2.6% somatic mutation, range 0%–38.8%).[13] Further analysis of our correlations showed that combining substrate scores improved predictions (Figures S7F and S7G). This average range (0.7%–27.5%) represents a cumulative ceiling (overfit) and floor (unoptimized) of our tool and suggests the need for proteomic measures to understand the consequences of mutational events better.

Although it may be difficult to predict individual proteins, consistent changes across many proteins in a pathway might reveal biologically meaningful results. To illustrate this, we aggregated C3PO scores across cancer hallmark genes[66,67] (Figures 7A, 7B, and S7I; Table S7; STAR Methods). C3PO was designed to assess variability between tumors. Hence, instances where nearly all tumors show pathway dysregulation, such as ~95% of PDACs being *KRAS* mutants, are not expected to be captured in this analysis. We explored hallmark variability between cancer types contributed by genomic variants by assessing entropy among the top three hallmark scores produced by C3PO for each sample (Figure 7C). From these predictions, we found certain cancer types showed lower hallmark variability (Figure 7D) than others (Figure 7E). For example, in LUAD, C3PO hallmark scores for adaptive immunity and DNA repair were consistently high, with enrichment of higher smoking scores[68] in the chromatin modification pathway (p value = 3.8×10^{-7}, chi-squared test) (Figure 7D; STAR Methods). In contrast, UCEC had substantially more heterogeneity for top hallmarks in each sample (Figure 7E), including DNA repair and pathways, such as NOTCH, not seen in LUAD. Thus, our current systematic understanding of genomics data can capture at least some variability in the protein levels of cancer hallmarks.

In addition to hallmark scores produced by C3PO, we used an established pathway enrichment tool, ssGSEA, to identify enrichment directly from protein abundance[69] (Figures 7A and 7F; Table S7; STAR Methods). We observed higher pathway variability in the top hallmarks per sample (Figure 7G). Similar hallmark uniformity was observed across all LUAD for DNA repair, cell cycle and apoptosis, and cancer drivers shared with C3PO, but many other hallmarks displayed much higher diversity (Figure S7H). Similarly, DNA repair, cell cycle and apoptosis, chromatin modification, and cancer driver hallmarks clustered together for C3PO and ssGSEA, but ssGSEA more diversely represented the remaining hallmarks. Upon further analysis, both LUAD (Figure 7H) and UCEC (Figure 7I) contained subgroups of samples with higher EMT and ECM activity—a hallmark not typically associated with known driver events (n = 19 and n = 21, respectively). The proteins most differentially expressed within these subgroups were VPS13A, FSTL1, and TMEM30B for UCEC (p values = 5.1×10^{-7}, 2.6×10^{-5}, and 6.3×10^{-5},

Figure 7. C3PO tool to investigate the multi-genomic impact on cancer endophenotypes

(A) Overview of C3PO. Left indicates the theoretical contribution of genomic alterations to protein levels (CNA and mutations); right provides a protein-only perspective of cancer hallmarks.

(B) Circos plot shows 21 cancer hallmark pathways. Icons indicate relationships to the original hallmarks outlined by Hanahan and Weinberg.[66,67] Every sample in CPTAC is represented by the links connecting hallmarks and is colored according to cancer type. Each sample is displayed using three lines according to their top three hallmark activity scores generated by C3PO. The histogram indicates the distribution of C3PO hallmark scores.

(C) Diversity among the top three hallmarks for each cancer type was quantified using Shannon entropy. Each bar shows the ranked hallmark entropy for the top three hallmarks in each sample and cancer type. Cancers are ordered by the lowest entropy measures of their sample's top hallmarks.

(D) Heatmap displays LUAD samples and their C3PO hallmark scores for each of the 21 hallmarks. Rows and columns are ordered using unsupervised clustering. Bottom annotations include neoantigen count estimation, total mutation counts, and smoking scores.

(E) Similar to Figure 7D, displays UCEC samples.

(F) Similar to Figure 7B, each sample is displayed with one line that links their top two hallmark scores produced by ssGSEA.

(G) Similar to Figure 7C, Shannon entropy calculations were made on ssGSEA scores.

(H–I) Similar to Figures 7D and 7E but show scores generated by ssGSEA.

See also Figure S7 and Table S7.

respectively; two-sided t test) and SLPI, TSPAN1, and CALD1 for LUAD (p values = 1.6×10^{-5}, 2.6×10^{-5}, and 4.4×10^{-5}, respectively; two-sided t test). This result, specific to our protein-only analysis from ssGSEA, implies that protein variability within tumors can reflect the aggregation of different alterations at the DNA, RNA, and protein levels. These can yield similar downstream effects and be independent of somatic alterations, possibly due to extracellular or microenvironment interactions. These findings complement genomics and highlight the proteomic contribution to tumor phenotype characterization.

DISCUSSION

Discovery and functional characterization of oncogenic drivers, accelerated by high-throughput genomic analyses, have advanced our mechanistic understanding of carcinogenesis, leading to more effective therapies.[25] Initially, targeted therapies were restricted to specific tumor types. However, the discovery of shared, therapeutically actionable drivers across different cancers has led to FDA drug approvals that are tumor-type agnostic. Subsequently, the FDA has approved the testing of broad gene panels to reveal clinically actionable variants,[70] where specific alterations are matched with targeted therapies.[71,72] These developments have spurred analyses of molecular underpinnings of pan-cancer drivers. TCGA pan-cancer atlas[73] represents an initial framework for integrating molecular genomic data from multiple tumor types, providing novel insights.[74–76] This also highlighted the need to expand the characterization of functional, mechanistic, and phenotypic correlates of driver alterations across additional layers of molecular data.

Here, we used proteomic and phosphoproteomic readouts from ten cancer types, integrated with genomic and transcriptomic data, to assess the pan-cancer consequences of 5,443 putative driver alterations across the CPTAC cohort. Our proteogenomic analyses provide insights into the molecular mechanisms of oncogenic mutations, from individual proteins to cancer hallmarks, revealing potentially novel therapeutic avenues. To precisely dissect how the cancer proteome is shaped, we analyzed the impact of oncogenic variants across different layers of biological regulation, starting with the effect on the same protein (*cis*-effect), moving to protein interactions and complexes, leading up to the entire (phospho-)proteome including *trans*-effects and polygenic predictive framework. To assess the rewiring of protein-protein interactions (PPI) in cancer,[40,77] we used proteomic co-expression data as an indirect readout of PPIs. We found candidate PPIs that differ by cancer type or driver alterations. Interestingly, many drivers were noted to display driver alterations at the interface between known interacting proteins, including PIK3R1-PIK3CA, SMAD4-SMAD2, and PPP2R1A-PPP2R2A. Importantly, similar analysis with the RNA data yielded only a subset of the PPI effects detected at the protein level, highlighting the importance of proteomic analysis. Given the emerging role of network medicine,[78] we trust that our use of pan-cancer proteomics data to associate mutations with the PPI network indirectly will be enlightening.

Our use of *trans*-effects of driver events revealed that different cancer genes in a pathway tend to display similar molecular fingerprints. This suggests that their molecular effects are similar, for example, *NFE2L2* and *KEAP1*, and *KMT2B* and *CREBBP*. Such molecular convergence explains the mutual exclusivity of a significant fraction of oncogenic drivers. However, we also found some cases where the molecular fingerprints are negatively correlated, and these often overlapped with mutually exclusive driver genes such as *EGFR* and *STK11*, *CDH1* and *TP53*, or *EGFR* and *KRAS*. It is conceivable that these pairs of genes are mutually exclusive because a mutation in one gene moves the cell toward a state where a mutation in the second gene is incompatible with oncogenesis (among other possible explanations). These genes (*EGFR* and *STK11*) are divergently incompatible rather than oncogenically convergent and redundant. These divergent incompatibilities could represent synthetic lethal vulnerabilities, as shown for *EGFR* and *KRAS* in lung cancer.[79] Drugs that exploit this phenomenon by activating genes divergently incompatible could be contemplated. Thus, the results from the Kinase Library, showing how *EGFR*-mutated and *STK11/KRAS*-mutated tumors activate opposite sets of kinases, could provide such drug targets.

To examine the cumulative impact of tumor genomic variants on the proteome, we built C3PO, a polygenic predictive framework trained to predict protein abundance. The predictive capacity is currently limited to ~27% of the global proteomic landscape, likely due to cellular plasticity[80] and transcriptomic fluctuations caused by the TME.[81] Nevertheless, this tool enables assessment of the genomic contribution to protein hallmark variability in tumors across space and time.[82]

In summary, this study highlights critical insights provided by proteomics to systematically assess the consequences of oncogenic drivers on the functional states of cancers. Going forward, a broader characterization of PTMs and the metabolome could further reveal how driver alterations perturb the activity of proteins such as E3 ubiquitin ligases. Furthermore, by applying single-cell proteomics,[83–85] spatial proteomics,[86] and multiple intra-/inter-tumor samples per patient, the proteomic contributions to tumor heterogeneity and interaction with the TME could be elucidated more comprehensively. Finally, clinical trials that couple proteomics to both pre- and post-treatment samples could reveal determinants of response and resistance to therapies and inform combination treatments at a level more directly related to the action of drugs: the proteome. This could lead to clinically implementable proteogenomic panels. Our findings support the proteome as a missing link between the genotype of oncogenic drivers and their functional states.

Limitations of the study

The pan-cancer proteogenomics study comprised ten tumor types previously analyzed individually as part of CPTAC consortium flagship studies.[13–22] This cohort is highly disparate, including three female-specific cancers, BRCA, HGSC, and UCEC, but not the most common male cancer, prostate adenocarcinoma. Also, GBM represents a relatively less common and biologically more distinct tumor type than the rest of the cohort. Additional planned cohorts will ameliorate these limitations of cohort composition. Some of the tumors in this study have cognate NAT composed of a cell lineage closely related to that of cancer. For others, this was unavailable. The pan-cancer analysis of drug sensitivity seems to focus on cell-cycle-related proteins, motivated in considerable measure by the interest in the community to target this core hallmark of cancer therapeutically. Although

finding that a kinase has elevated activity in tumors might suggest a potential drug target, careful consideration of the essentiality of that kinase in normal tissues is needed to establish a possible therapeutic window. Nevertheless, a pan-cancer proteomic readout of other hallmark phenotypes may yet be therapeutically informative and would merit in-depth assessment in future studies.

CONSORTIA

The members of the National Cancer Institute Clinical Proteomic Tumor Analysis Consortium for pan-cancer are François Aguet, Yo Akiyama, Eunkyung An, Shankara Anand, Meenakshi Anurag, Özgün Babur, Jasmin Bavarva, Chet Birger, Michael J. Birrer, Anna Calinawan, Lewis C. Cantley, Song Cao, Steven A. Carr, Michele Ceccarelli, Daniel W. Chan, Arul M. Chinnaiyan, Hanbyul Cho, Shrabanti Chowdhury, Marcin P. Cieslik, Karl R. Clauser, Antonio Colaprico, Daniel Cui Zhou, Felipe da Veiga Leprevost, Corbin Day, Saravana M. Dhanasekaran, Li Ding, Marcin J. Domagalski, Yongchao Dou, Brian J. Druker, Nathan Edwards, Matthew J. Ellis, Myvizhi Esai Selvan, David Fenyö, Steven M. Foltz, Alicia Francis, Yifat Geffen, Gad Getz, Michael A. Gillette, Tania J. Gonzalez Robles, Sara J.C. Gosline, Zeynep H. Gümüş, David I. Heiman, Tara Hiltke, Runyu Hong, Galen Hostetter, Yingwei Hu, Chen Huang, Emily Huntsman, Antonio Iavarone, Eric J. Jaehnig, Scott D. Jewell, Jiayi Ji, Wen Jiang, Jared L. Johnson, Lizabeth Katsnelson, Karen A. Ketchum, Iga Kolodziejczak, Karsten Krug, Chandan Kumar-Sinha, Alexander J. Lazar, Jonathan T. Lei, Yize Li, Wen-Wei Liang, Yuxing Liao, Caleb M. Lindgren, Tao Liu, Wenke Liu, Weiping Ma, D.R. Mani, Fernanda Martins Rodrigues, Wilson McKerrow, Mehdi Mesri, Alexey I. Nesvizhskii, Chelsea J. Newton, Robert Oldroyd, Gilbert S. Omenn, Amanda G. Paulovich, Samuel H. Payne, Francesca Petralia, Pietro Pugliese, Boris Reva, Ana I. Robles, Karin D. Rodland, Henry Rodriguez, Kelly V. Ruggles, Dmitry Rykunov, Shankha Satpathy, Sara R. Savage, Eric E. Schadt, Michael Schnaubelt, Tobias Schraink, Stephan Schürer, Zhiao Shi, Richard D. Smith, Xiaoyu Song, Yizhe Song, Vasileios Stathias, Erik P. Storrs, Jimin Tan, Nadezhda V. Terekhanova, Ratna R. Thangudu, Mathangi Thiagarajan, Nicole Tignor, Joshua M. Wang, Liang-Bo Wang, Pei Wang, Ying Wang, Bo Wen, Maciej Wiznerowicz, Yige Wu, Matthew A. Wyczalkowski, Lijun Yao, Tomer M. Yaron, Xinpei Yi, Bing Zhang, Hui Zhang, Qing Zhang, Xu Zhang, and Zhen Zhang.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.cell.2023.07.014.

ACKNOWLEDGMENTS

This work was supported by the National Cancer Institute (NCI) Clinical Proteomic Tumor Analysis Consortium (CPTAC) under award numbers U24CA210955, U24CA210985, U24CA210986, U24CA210954, U24CA210967, U24CA210972,

U24CA210979, U24CA210993, U01CA214114, U01CA214116, and U01CA214125 as U24CA210972 (D.F., L.D., and S.H.P.), U24CA210979 (G.G. and C.B.), and U24CA270823 (S.S. and M.A.G.), and Contract GR0012005 (L.D.). M.S. is supported by the Asociación Española Contra el Cáncer (AECC) project LABAE20038PORT. E.P.-P., V.R.-S., and K.J.I. are supported by the Spanish Ministry of Science (grants RYC2019-026415-I and PID2019-107043RA-I00).

AUTHOR CONTRIBUTIONS

Study conception and design, Y.L., C.K.-S., A.J.L., L.C.C., G.G., and L.D.; performed experiment or data collection, Y.L., Y.G., S.A., M.A.W., Y.S., W.-W.L., N.V.T., F.M.R., K.R.C., D.I.H., Q.Z., F.A., C.B., D.C.Z., and L.-B.W.; computational, multi-omic, and statistical analyses, Y.L., E.P.-P., C.T., M.H.B., T.M.Y., V.S., K.J.I., Y.G., S.C., S.A., Y.A., W.L., E.P.S., M.C.W., W.Z., M.S., V.R.-S., J.L.J., E.M.H., P.P., and A.J.L.; data interpretation and biological analysis, Y.L., E.P.-P., C.T., M.H.B., T.M.Y., Y.G., V.S., K.J.I., S.C., S.A., Y.A., W.L., E.P.S., M.C.W., W.Z., M.S., V.R.-S., J.L.J., E.M.H., P.P., C.K.-S., A.J.L., L.C.C., G.G., and L.D.; data visualization and figure design: Y.L., E.P-P., C.T., M.H.B., T.M.Y., V.S., K.J.I., S.C., M.A.W., C.K.-S., A.J.L., L.C.C., G.G., and L.D; writing – original drafts, Y.L., E.P.-P., C.T., M.H.B., T.M.Y., V.S., K.J.I., S.C., C.K.-S., A.J.L., and L.D.; writing – review and editing, Y.L., E.P.-P., C.T., M.H.B., T.M.Y., V.S., Y.G., K.J.I., S.C., M.C.W., S.M.D., S.S., J.B., A.C., A.I., M.C., C.J.R., D.F., S.H.P., A.I.R., M.A.G., C.K.-S., A.J.L., L.C.C., G.G., and L.D.; supervision: Y.L., A.I.R., C.K.-S., A.J.L., L.C.C., G.G., and L.D.; and administration, Y.L., H.R., A.I.R., G.G., and L.D.

DECLARATION OF INTERESTS

Y.G. is a consultant for Oriel Research Therapeutics. T.M.Y. is a co-founder, stockholder, and member of the board of directors of DESTROKE, Inc., an early-stage start-up developing mobile technology for automated clinical stroke detection. J.L.J. has received consulting fees from Scorpion Therapeutics and Volastra Therapeutics. F.A. is an inventor on a patent application related to SignatureAnalyzer-GPU and has been an employee of Illumina, Inc., since 8 November 2021. L.C.C. is a founder and member of the board of directors of Agios Pharmaceuticals and is a founder of Petra Pharmaceuticals. L.C.C. is an inventor on patents (pending) for Combination Therapy for PI3K-associated Disease or Disorder, and The Identification of Therapeutic Interventions to Improve Response to PI3K Inhibitors for Cancer Treatment. L.C.C. is a co-founder and shareholder in Faeth Therapeutics. G.G. receives research funds from IBM, Pharmacyclics, and Ultima Genomics and is also an inventor on patent applications filed by the Broad Institute related to MSMuTect, MSMutSig, POLYSOLVER, SignatureAnalyzer-GPU, MSIDetect, and MinimuMM-Seq. G.G. is also a founder, consultant, and privately held equity in Scorpion Therapeutics.

Received: July 5, 2022
Revised: December 30, 2022
Accepted: July 10, 2023
Published: August 14, 2023

REFERENCES

1. Bailey, M.H., Tokheim, C., Porta-Pardo, E., Sengupta, S., Bertrand, D., Weerasinghe, A., Colaprico, A., Wendl, M.C., Kim, J., Reardon, B., et al. (2018). Comprehensive characterization of cancer driver genes and mutations. Cell *173*, 371–385.e18. https://doi.org/10.1016/j.cell.2018.02.060.
2. Lawrence, M.S., Stojanov, P., Polak, P., Kryukov, G.V., Cibulskis, K., Sivachenko, A., Carter, S.L., Stewart, C., Mermel, C.H., Roberts, S.A., et al. (2013). Mutational heterogeneity in cancer and the search for new cancer-associated genes. Nature *499*, 214–218. https://doi.org/10.1038/nature12213.
3. Porta-Pardo, E., and Godzik, A. (2014). e-Driver: a novel method to identify protein regions driving cancer. Bioinformatics *30*, 3109–3114. https://doi.org/10.1093/bioinformatics/btu499.
4. Tokheim, C., and Karchin, R. (2019). CHASMplus reveals the scope of somatic missense mutations driving human cancers. Cell Syst. *9*, 9–23.e8. https://doi.org/10.1016/j.cels.2019.05.005.
5. Muiños, F., Martínez-Jiménez, F., Pich, O., Gonzalez-Perez, A., and Lopez-Bigas, N. (2021). In silico saturation mutagenesis of cancer genes. Nature *596*, 428–432. https://doi.org/10.1038/s41586-021-03771-1.
6. Hess, J.M., Bernards, A., Kim, J., Miller, M., Taylor-Weiner, A., Haradhvala, N.J., Lawrence, M.S., and Getz, G. (2019). Passenger hotspot mutations in cancer. Cancer Cell *36*, 288–301.e14. https://doi.org/10.1016/j.ccell.2019.08.002.
7. Mularoni, L., Sabarinathan, R., Deu-Pons, J., Gonzalez-Perez, A., and López-Bigas, N. (2016). OncodriveFML: A general framework to identify coding and non-coding regions with cancer driver mutations. Genome Biol. *17*, 128. https://doi.org/10.1186/s13059-016-0994-0.
8. Kamburov, A., Lawrence, M.S., Polak, P., Leshchiner, I., Lage, K., Golub, T.R., Lander, E.S., and Getz, G. (2015). Comprehensive assessment of cancer missense mutation clustering in protein structures. Proc. Natl. Acad. Sci. USA. *112*, E5486–E5495. https://doi.org/10.1073/pnas.1516373112.
9. Porta-Pardo, E., Garcia-Alonso, L., Hrabe, T., Dopazo, J., and Godzik, A. (2015). A pan-cancer catalogue of cancer driver protein interaction interfaces. PLoS Comput. Biol. *11*, e1004518. https://doi.org/10.1371/journal.pcbi.1004518.
10. Tokheim, C., Bhattacharya, R., Niknafs, N., Gygax, D.M., Kim, R., Ryan, M., Masica, D.L., and Karchin, R. (2016). Exome-scale discovery of hotspot mutation regions in human cancer using 3D protein structure. Cancer Res. *76*, 3719–3731. https://doi.org/10.1158/0008-5472.CAN-15-3190.
11. Kandoth, C., McLellan, M.D., Vandin, F., Ye, K., Niu, B., Lu, C., Xie, M., Zhang, Q., McMichael, J.F., Wyczalkowski, M.A., et al. (2013). Mutational landscape and significance across 12 major cancer types. Nature *502*, 333–339. https://doi.org/10.1038/nature12634.
12. Grandér, D. (1998). How do mutated oncogenes and tumor suppressor genes cause cancer? Med. Oncol. *15*, 20–26. https://doi.org/10.1007/BF02787340.
13. Vasaikar, S., Huang, C., Wang, X., Petyuk, V.A., Savage, S.R., Wen, B., Dou, Y., Zhang, Y., Shi, Z., Arshad, O.A., et al. (2019). Proteogenomic analysis of human colon cancer reveals new therapeutic opportunities. Cell *177*, 1035–1049.e19. https://doi.org/10.1016/j.cell.2019.03.030.
14. Clark, D.J., Dhanasekaran, S.M., Petralia, F., Pan, J., Song, X., Hu, Y., da Veiga Leprevost, F., Reva, B., Lih, T.M., Chang, H.Y., et al. (2019). Integrated proteogenomic characterization of clear cell renal cell carcinoma. Cell *179*, 964–983.e31. https://doi.org/10.1016/j.cell.2019.10.007.
15. Dou, Y., Kawaler, E.A., Cui Zhou, D., Gritsenko, M.A., Huang, C., Blumenberg, L., Karpova, A., Petyuk, V.A., Savage, S.R., Satpathy, S., et al. (2020). Proteogenomic characterization of endometrial carcinoma. Cell *180*, 729–748.e26. https://doi.org/10.1016/j.cell.2020.01.026.
16. Wang, L.B., Karpova, A., Gritsenko, M.A., Kyle, J.E., Cao, S., Li, Y., Rykunov, D., Colaprico, A., Rothstein, J.H., Hong, R., et al. (2021). Proteogenomic and metabolomic characterization of human glioblastoma. Cancer Cell *39*, 509–528.e20. https://doi.org/10.1016/j.ccell.2021.01.006.
17. Gillette, M.A., Satpathy, S., Cao, S., Dhanasekaran, S.M., Vasaikar, S.V., Krug, K., Petralia, F., Li, Y., Liang, W.W., Reva, B., et al. (2020). Proteogenomic characterization reveals therapeutic vulnerabilities in lung adenocarcinoma. Cell *182*, 200–225.e35. https://doi.org/10.1016/j.cell.2020.06.013.
18. Huang, C., Chen, L., Savage, S.R., Eguez, R.V., Dou, Y., Li, Y., da Veiga Leprevost, F., Jaehnig, E.J., Lei, J.T., Wen, B., et al. (2021). Proteogenomic insights into the biology and treatment of HPV-negative head

and neck squamous cell carcinoma. Cancer Cell *39*, 361–379.e16. https://doi.org/10.1016/j.ccell.2020.12.007.

19. Krug, K., Jaehnig, E.J., Satpathy, S., Blumenberg, L., Karpova, A., Anurag, M., Miles, G., Mertins, P., Geffen, Y., Tang, L.C., et al. (2020). Proteogenomic landscape of breast cancer tumorigenesis and targeted therapy. Cell *183*, 1436–1456.e31. https://doi.org/10.1016/j.cell.2020.10.036.
20. McDermott, J.E., Arshad, O.A., Petyuk, V.A., Fu, Y., Gritsenko, M.A., Clauss, T.R., Moore, R.J., Schepmoes, A.A., Zhao, R., Monroe, M.E., et al. (2020). Proteogenomic characterization of ovarian HGSC implicates mitotic kinases, replication stress in observed chromosomal instability. Cell Rep. Med. *1*, 100004. https://doi.org/10.1016/j.xcrm.2020.100004.
21. Cao, L., Huang, C., Cui Zhou, D., Hu, Y., Lih, T.M., Savage, S.R., Krug, K., Clark, D.J., Schnaubelt, M., Chen, L., et al. (2021). Proteogenomic characterization of pancreatic ductal adenocarcinoma. Cell *184*, 5031–5052.e26. https://doi.org/10.1016/j.cell.2021.08.023.
22. Satpathy, S., Krug, K., Jean Beltran, P.M., Savage, S.R., Petralia, F., Kumar-Sinha, C., Dou, Y., Reva, B., Kane, M.H., Avanessian, S.C., et al. (2021). A proteogenomic portrait of lung squamous cell carcinoma. Cell *184*, 4348–4371.e40. https://doi.org/10.1016/j.cell.2021.07.016.
23. Cao, Y., Zhou, W., Li, L., Wang, J., Gao, Z., Jiang, Y., Jiang, X., Shan, A., Bailey, M.H., Huang, K.L., et al. (2018). Pan-cancer analysis of somatic mutations across 21 neuroendocrine tumor types. Cell Res. *28*, 601–604. https://doi.org/10.1038/s41422-018-0019-5.
24. ICGC/TCGA Pan-Cancer Analysis of Whole Genomes (2020). Pan-cancer analysis of whole genomes. Nature *578*, 82–93. https://doi.org/10.1038/s41586-020-1969-6.
25. Chen, F., Wendl, M.C., Wyczalkowski, M.A., Bailey, M.H., Li, Y., and Ding, L. (2021). Moving pan-cancer studies from basic research toward the clinic. Nat. Cancer *2*, 879–890. https://doi.org/10.1038/s43018-021-00250-4.
26. Bailey, M.H., Meyerson, W.U., Dursi, L.J., Wang, L.B., Dong, G., Liang, W.W., Weerasinghe, A., Li, S., Li, Y., Kelso, S., et al. (2020). Retrospective evaluation of whole exome and genome mutation calls in 746 cancer samples. Nat. Commun. *11*, 4748. https://doi.org/10.1038/s41467-020-18151-y.
27. Akbani, R., Ng, P.K., Werner, H.M., Shahmoradgoli, M., Zhang, F., Ju, Z., Liu, W., Yang, J.Y., Yoshihara, K., Li, J., et al. (2014). A pan-cancer proteomic perspective on The Cancer Genome Atlas. Nat. Commun. *5*, 3887. https://doi.org/10.1038/ncomms4887.
28. Kim, J., Mouw, K.W., Polak, P., Braunstein, L.Z., Kamburov, A., Kwiatkowski, D.J., Rosenberg, J.E., Van Allen, E.M., D'Andrea, A., and Getz, G. (2016). Somatic ERCC2 mutations are associated with a distinct genomic signature in urothelial tumors. Nat. Genet. *48*, 600–606. https://doi.org/10.1038/ng.3557.
29. Kasar, S., Kim, J., Improgo, R., Tiao, G., Polak, P., Haradhvala, N., Lawrence, M.S., Kiezun, A., Fernandes, S.M., Bahl, S., et al. (2015). Whole-genome sequencing reveals activation-induced cytidine deaminase signatures during indolent chronic lymphocytic leukaemia evolution. Nat. Commun. *6*, 8866. https://doi.org/10.1038/ncomms9866.
30. Taylor-Weiner, A., Aguet, F., Haradhvala, N.J., Gosai, S., Anand, S., Kim, J., Ardlie, K., Van Allen, E.M., and Getz, G. (2019). Scaling computational genomics to millions of individuals with GPUs. Genome Biol. *20*, 228. https://doi.org/10.1186/s13059-019-1836-7.
31. Thorsson, V., Gibbs, D.L., Brown, S.D., Wolf, D., Bortone, D.S., Ou Yang, T.H., Porta-Pardo, E., Gao, G.F., Plaisier, C.L., Eddy, J.A., et al. (2018). The immune landscape of cancer. Immunity *48*, 812–830.e14. https://doi.org/10.1016/j.immuni.2018.03.023.
32. Phillips, H.S., Kharbanda, S., Chen, R., Forrest, W.F., Soriano, R.H., Wu, T.D., Misra, A., Nigro, J.M., Colman, H., Soroceanu, L., et al. (2006). Molecular subclasses of high-grade glioma predict prognosis, delineate a pattern of disease progression, and resemble stages in neurogenesis. Cancer Cell *9*, 157–173. https://doi.org/10.1016/j.ccr.2006.02.019.
33. Verhaak, R.G., Hoadley, K.A., Purdom, E., Wang, V., Qi, Y., Wilkerson, M.D., Miller, C.R., Ding, L., Golub, T., Mesirov, J.P., et al. (2010). Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in PDGFRA, IDH1, EGFR, and NF1. Cancer Cell *17*, 98–110. https://doi.org/10.1016/j.ccr.2009.12.020.
34. Zarkoob, H., Taube, J.H., Singh, S.K., Mani, S.A., and Kohandel, M. (2013). Investigating the link between molecular subtypes of glioblastoma, epithelial-mesenchymal transition, and CD133 cell surface protein. PLoS One *8*, e64169. https://doi.org/10.1371/journal.pone.0064169.
35. Pietzner, M., Wheeler, E., Carrasco-Zanini, J., Kerrison, N.D., Oerton, E., Koprulu, M., Luan, J., Hingorani, A.D., Williams, S.A., Wareham, N.J., and Langenberg, C. (2021). Synergistic insights into human health from aptamer- and antibody-based proteomic profiling. Nat. Commun. *12*, 6822. https://doi.org/10.1038/s41467-021-27164-0.
36. Chick, J.M., Munger, S.C., Simecek, P., Huttlin, E.L., Choi, K., Gatti, D.M., Raghupathy, N., Svenson, K.L., Churchill, G.A., and Gygi, S.P. (2016). Defining the consequences of genetic variation on a proteome-wide scale. Nature *534*, 500–505. https://doi.org/10.1038/nature18270.
37. Freed-Pastor, W.A., and Prives, C. (2012). Mutant p53: one name, many proteins. Genes Dev. *26*, 1268–1286. https://doi.org/10.1101/gad.190678.112.
38. Mighell, T.L., Evans-Dutson, S., and O'Roak, B.J. (2018). A saturation mutagenesis approach to understanding PTEN lipid phosphatase activity and genotype-phenotype relationships. Am. J. Hum. Genet. *102*, 943–955. https://doi.org/10.1016/j.ajhg.2018.03.018.
39. Matreyek, K.A., Starita, L.M., Stephany, J.J., Martin, B., Chiasson, M.A., Gray, V.E., Kircher, M., Khechaduri, A., Dines, J.N., Hause, R.J., et al. (2018). Multiplex assessment of protein variant abundance by massively parallel sequencing. Nat. Genet. *50*, 874–882. https://doi.org/10.1038/s41588-018-0122-z.
40. Mo, X., Niu, Q., Ivanov, A.A., Tsang, Y.H., Tang, C., Shu, C., Li, Q., Qian, K., Wahafu, A., Doyle, S.P., et al. (2022). Systematic discovery of mutation-directed neo-protein-protein interactions in cancer. Cell *185*, 1974–1985.e12. https://doi.org/10.1016/j.cell.2022.04.014.
41. Romanov, N., Kuhn, M., Aebersold, R., Ori, A., Beck, M., and Bork, P. (2019). Disentangling genetic and environmental effects on the proteotypes of individuals. Cell *177*, 1308–1318.e10. https://doi.org/10.1016/j.cell.2019.03.015.
42. Sarbassov, D.D., Ali, S.M., Kim, D.H., Guertin, D.A., Latek, R.R., Erdjument-Bromage, H., Tempst, P., and Sabatini, D.M. (2004). Rictor, a novel binding partner of mTOR, defines a rapamycin-insensitive and raptor-independent pathway that regulates the cytoskeleton. Curr. Biol. *14*, 1296–1302. https://doi.org/10.1016/j.cub.2004.06.054.
43. van der Wal, T., and van Amerongen, R. (2020). Walking the tight wire between cell adhesion and WNT signalling: a balancing act for beta-catenin. Open Biol. *10*, 200267. https://doi.org/10.1098/rsob.200267.
44. Schick, S., Rendeiro, A.F., Runggatscher, K., Ringler, A., Boidol, B., Hinkel, M., Májek, P., Vulliard, L., Penz, T., Parapatics, K., et al. (2019). Systematic characterization of BAF mutations provides insights into intracomplex synthetic lethalities in human cancers. Nat. Genet. *51*, 1399–1410. https://doi.org/10.1038/s41588-019-0477-9.
45. Kluba, M., Engelborghs, Y., Hofkens, J., and Mizuno, H. (2015). Inhibition of receptor dimerization as a novel negative feedback mechanism of EGFR signaling. PLoS One *10*, e0139971. https://doi.org/10.1371/journal.pone.0139971.
46. Thirukkumaran, O.M., Kluba, M., Hofkens, J., and Mizuno, H. (2020). Autophosphorylation of EGFR at Y954 facilitated homodimerization and enhanced downstream signals. Biophys. J. *119*, 2127–2137. https://doi.org/10.1016/j.bpj.2020.10.008.
47. McMahon, M., Itoh, K., Yamamoto, M., and Hayes, J.D. (2003). Keap1-dependent proteasomal degradation of transcription factor Nrf2 contributes to the negative regulation of antioxidant response element-driven gene expression. J. Biol. Chem. *278*, 21592–21600. https://doi.org/10.1074/jbc.M300931200.

48. Wang, M.T., Holderfield, M., Galeas, J., Delrosario, R., To, M.D., Balmain, A., and McCormick, F. (2015). K-Ras promotes tumorigenicity through suppression of non-canonical Wnt signaling. Cell *163*, 1237–1251. https://doi.org/10.1016/j.cell.2015.10.041.

49. Bokhari, A., Jonchere, V., Lagrange, A., Bertrand, R., Svrcek, M., Marisa, L., Buhard, O., Greene, M., Demidova, A., Jia, J., et al. (2018). Targeting nonsense-mediated mRNA decay in colorectal cancers with microsatellite instability. Oncogenesis *7*, 70. https://doi.org/10.1038/s41389-018-0079-x.

50. Lake, D., Corrêa, S.A., and Müller, J. (2016). Negative feedback regulation of the ERK1/2 MAPK pathway. Cell. Mol. Life Sci. *73*, 4397–4413. https://doi.org/10.1007/s00018-016-2297-8.

51. Schumacher, T.N., and Schreiber, R.D. (2015). Neoantigens in cancer immunotherapy. Science *348*, 69–74. https://doi.org/10.1126/science.aaa4971.

52. Turajlic, S., Litchfield, K., Xu, H., Rosenthal, R., McGranahan, N., Reading, J.L., Wong, Y.N.S., Rowan, A., Kanu, N., Al Bakir, M., et al. (2017). Insertion-and-deletion-derived tumour-specific neoantigens and the immunogenic phenotype: a pan-cancer analysis. Lancet Oncol. *18*, 1009–1021. https://doi.org/10.1016/S1470-2045(17)30516-8.

53. Young, R.J., Waldeck, K., Martin, C., Foo, J.H., Cameron, D.P., Kirby, L., Do, H., Mitchell, C., Cullinane, C., Liu, W., et al. (2014). Loss of CDKN2A expression is a frequent event in primary invasive melanoma and correlates with sensitivity to the CDK4/6 inhibitor PD0332991 in melanoma cell lines. Pigment Cell Melanoma Res. *27*, 590–600. https://doi.org/10.1111/pcmr.12228.

54. Elvin, J.A., Gay, L.M., Ort, R., Shuluk, J., Long, J., Shelley, L., Lee, R., Chalmers, Z.R., Frampton, G.M., Ali, S.M., et al. (2017). Clinical benefit in response to palbociclib treatment in refractory uterine leiomyosarcomas with a common CDKN2A alteration. Oncologist *22*, 416–421. https://doi.org/10.1634/theoncologist.2016-0310.

55. Finn, R.S., Dering, J., Conklin, D., Kalous, O., Cohen, D.J., Desai, A.J., Ginther, C., Atefi, M., Chen, I., Fowst, C., et al. (2009). PD 0332991, a selective cyclin D kinase 4/6 inhibitor, preferentially inhibits proliferation of luminal estrogen receptor-positive human breast cancer cell lines in vitro. Breast Cancer Res. *11*, R77. https://doi.org/10.1186/bcr2419.

56. Dyson, N.J. (2016). RB1: a prototype tumor suppressor and an enigma. Genes Dev. *30*, 1492–1502. https://doi.org/10.1101/gad.282145.116.

57. Subramanian, A., Narayan, R., Corsello, S.M., Peck, D.D., Natoli, T.E., Lu, X., Gould, J., Davis, J.F., Tubelli, A.A., Asiedu, J.K., et al. (2017). A next generation connectivity map: L1000 platform and the first 1,000,000 profiles. Cell *171*, 1437–1452.e17. https://doi.org/10.1016/j.cell.2017.10.049.

58. Condorelli, R., Spring, L., O'Shaughnessy, J., Lacroix, L., Bailleux, C., Scott, V., Dubois, J., Nagy, R.J., Lanman, R.B., Iafrate, A.J., et al. (2018). Polyclonal RB1 mutations and acquired resistance to CDK 4/6 inhibitors in patients with metastatic breast cancer. Ann. Oncol. *29*, 640–645. https://doi.org/10.1093/annonc/mdx784.

59. Wang, B., Li, R., Wu, S., Liu, X., Ren, J., Li, J., Bi, K., Wang, Y., and Jia, H. (2021). Breast cancer resistance to cyclin-dependent kinases 4/6 inhibitors: intricacy of the molecular mechanisms. Front. Oncol. *11*, 651541. https://doi.org/10.3389/fonc.2021.651541.

60. Tsherniak, A., Vazquez, F., Montgomery, P.G., Weir, B.A., Kryukov, G., Cowley, G.S., Gill, S., Harrington, W.F., Pantel, S., Krill-Burger, J.M., et al. (2017). Defining a cancer dependency map. Cell *170*, 564–576.e16. https://doi.org/10.1016/j.cell.2017.06.010.

61. Wang, L., Shao, X., Zhong, T., Wu, Y., Xu, A., Sun, X., Gao, H., Liu, Y., Lan, T., Tong, Y., et al. (2021). Discovery of a first-in-class CDK2 selective degrader for AML differentiation therapy. Nat. Chem. Biol. *17*, 567–575. https://doi.org/10.1038/s41589-021-00742-5.

62. Torkamani, A., Wineinger, N.E., and Topol, E.J. (2018). The personal and clinical utility of polygenic risk scores. Nat. Rev. Genet. *19*, 581–590. https://doi.org/10.1038/s41576-018-0018-x.

63. Dudbridge, F. (2013). Power and predictive accuracy of polygenic risk scores. PLoS Genet. *9*, e1003348. https://doi.org/10.1371/journal.pgen.1003348.

64. Lewis, C.M., and Vassos, E. (2020). Polygenic risk scores: from research tools to clinical instruments. Genome Med. *12*, 44. https://doi.org/10.1186/s13073-020-00742-5.

65. Zheng, F., Kelly, M.R., Ramms, D.J., Heintschel, M.L., Tao, K., Tutuncuoglu, B., Lee, J.J., Ono, K., Foussard, H., Chen, M., et al. (2021). Interpretation of cancer mutations using a multiscale map of protein systems. Science *374*, eabf3067. https://doi.org/10.1126/science.abf3067.

66. Hanahan, D., and Weinberg, R.A. (2011). Hallmarks of cancer: the next generation. Cell *144*, 646–674. https://doi.org/10.1016/j.cell.2011.02.013.

67. Hanahan, D. (2022). Hallmarks of cancer: new dimensions. Cancer Discov. *12*, 31–46. https://doi.org/10.1158/2159-8290.CD-21-1059.

68. Devarakonda, S., Li, Y., Martins Rodrigues, F., Sankararaman, S., Kadara, H., Goparaju, C., Lanc, I., Pepin, K., Waqar, S.N., Morgensztern, D., et al. (2021). Genomic profiling of lung adenocarcinoma in never-smokers. J. Clin. Oncol. *39*, 3747–3758. https://doi.org/10.1200/JCO.21.01691.

69. Barbie, D.A., Tamayo, P., Boehm, J.S., Kim, S.Y., Moody, S.E., Dunn, I.F., Schinzel, A.C., Sandy, P., Meylan, E., Scholl, C., et al. (2009). Systematic RNA interference reveals that oncogenic KRAS-driven cancers require TBK1. Nature *462*, 108–112. https://doi.org/10.1038/nature08460.

70. Cheng, D.T., Mitchell, T.N., Zehir, A., Shah, R.H., Benayed, R., Syed, A., Chandramohan, R., Liu, Z.Y., Won, H.H., Scott, S.N., et al. (2015). Memorial Sloan Kettering-integrated mutation profiling of actionable cancer targets (MSK-Impact): a hybridization capture-based next-generation sequencing clinical assay for solid tumor molecular oncology. J. Mol. Diagn. *17*, 251–264. https://doi.org/10.1016/j.jmoldx.2014.12.006.

71. Chakravarty, D., Gao, J., Phillips, S.M., Kundra, R., Zhang, H., Wang, J., Rudolph, J.E., Yaeger, R., Soumerai, T., Nissan, M.H., et al. (2017). OncoKB: a precision oncology knowledge base. JCO Precis. Oncol. *2017*. https://doi.org/10.1200/PO.17.00011.

72. Frampton, G.M., Fichtenholtz, A., Otto, G.A., Wang, K., Downing, S.R., He, J., Schnall-Levin, M., White, J., Sanford, E.M., An, P., et al. (2013). Development and validation of a clinical cancer genomic profiling test based on massively parallel DNA sequencing. Nat. Biotechnol. *31*, 1023–1031. https://doi.org/10.1038/nbt.2696.

73. Ding, L., Bailey, M.H., Porta-Pardo, E., Thorsson, V., Colaprico, A., Bertrand, D., Gibbs, D.L., Weerasinghe, A., Huang, K.L., Tokheim, C., et al. (2018). Perspective on oncogenic processes at the end of the beginning of cancer genomics. Cell *173*, 305–320.e10. https://doi.org/10.1016/j.cell.2018.03.033.

74. Cobain, E.F., Wu, Y.M., Vats, P., Chugh, R., Worden, F., Smith, D.C., Schuetze, S.M., Zalupski, M.M., Sahai, V., Alva, A., et al. (2021). Assessment of clinical benefit of integrative genomic profiling in advanced solid tumors. JAMA Oncol. *7*, 525–533. https://doi.org/10.1001/jamaoncol.2020.7987.

75. Marquart, J., Chen, E.Y., and Prasad, V. (2018). Estimation of the percentage of US patients with cancer who benefit from genome-driven oncology. JAMA Oncol. *4*, 1093–1098. https://doi.org/10.1001/jamaoncol.2018.1660.

76. Presley, C.J., Tang, D., Soulos, P.R., Chiang, A.C., Longtine, J.A., Adelson, K.B., Herbst, R.S., Zhu, W., Nussbaum, N.C., Sorg, R.A., et al. (2018). Association of broad-based genomic sequencing with survival among patients with advanced non-small cell lung cancer in the community oncology setting. JAMA *320*, 469–477. https://doi.org/10.1001/jama.2018.9824.

77. Kim, M., Park, J., Bouhaddou, M., Kim, K., Rojc, A., Modak, M., Soucheray, M., McGregor, M.J., O'Leary, P., Wolf, D., et al. (2021). A protein interaction landscape of breast cancer. Science *374*, eabf3066. https://doi.org/10.1126/science.abf3066.

78. Barabási, A.L., Gulbahce, N., and Loscalzo, J. (2011). Network medicine: a network-based approach to human disease. Nat. Rev. Genet. *12*, 56–68. https://doi.org/10.1038/nrg2918.

79. Unni, A.M., Lockwood, W.W., Zejnullahu, K., Lee-Lin, S.Q., and Varmus, H. (2015). Evidence that synthetic lethality underlies the mutual exclusivity of oncogenic KRAS and EGFR mutations in lung adenocarcinoma. eLife *4*, e06907. https://doi.org/10.7554/eLife.06907.

80. Meacham, C.E., and Morrison, S.J. (2013). Tumour heterogeneity and cancer cell plasticity. Nature *501*, 328–337. https://doi.org/10.1038/nature12624.

81. Barkley, D., Moncada, R., Pour, M., Liberman, D.A., Dryg, I., Werba, G., Wang, W., Baron, M., Rao, A., Xia, B., et al. (2022). Cancer cell states recur across tumor types and form specific interactions with the tumor microenvironment. Nat. Genet. *54*, 1192–1201. https://doi.org/10.1038/s41588-022-01141-9.

82. Sibai, M., Cervilla, S., Grases, D., Musulen, E., Fortian, A., Romeo, M., Bernat, A., Tokheim, C., Esteller, M., Barretina, J., et al. (2022). Charting the spatial landscape of cancer hallmarks. bioRxiv.. https://doi.org/10.1101/2022.06.18.496114.

83. Brunner, A.D., Thielert, M., Vasilopoulou, C., Ammar, C., Coscia, F., Mund, A., Hoerning, O.B., Bache, N., Apalategui, A., Lubeck, M., et al. (2022). Ultra-high sensitivity mass spectrometry quantifies single-cell proteome changes upon perturbation. Mol. Syst. Biol. *18*, e10798. https://doi.org/10.15252/msb.202110798.

84. Petelski, A.A., Emmott, E., Leduc, A., Huffman, R.G., Specht, H., Perlman, D.H., and Slavov, N. (2021). Multiplexed single-cell proteomics using SCoPE2. Nat. Protoc. *16*, 5398–5425. https://doi.org/10.1038/s41596-021-00616-z.

85. Slavov, N. (2022). Scaling up single-cell proteomics. Mol. Cell. Proteomics *21*, 100179. https://doi.org/10.1016/j.mcpro.2021.100179.

86. Lundberg, E., and Borner, G.H.H. (2019). Spatial proteomics: a powerful discovery tool for cell biology. Nat. Rev. Mol. Cell Biol. *20*, 285–302. https://doi.org/10.1038/s41580-018-0094-y.

87. Li, Y., Dou, Y., Da Veiga Leprevost, F., Geffen, Y., Calinawan, A.P., Aguet, F., Akiyama, Y., Anand, S., Birger, C., Cao, S., et al. (2023). Proteogenomic data and resources for pan-cancer analysis. Cancer Cell *41*, 1397–1406.

88. Li, Y., Lih, T.M., Dhanasekaran, S.M., Mannan, R., Chen, L., Cieslik, M., Wu, Y., Lu, R.J., Clark, D.J., Kołodziejczak, I., et al. (2023). Histopathologic and proteogenomic heterogeneity reveals features of clear cell renal cell carcinoma aggressiveness. Cancer Cell *41*, 139–163.e17. https://doi.org/10.1016/j.ccell.2022.12.001.

89. Cui Zhou, D., Jayasinghe, R.G., Chen, S., Herndon, J.M., Iglesia, M.D., Navale, P., Wendl, M.C., Caravan, W., Sato, K., Storrs, E., et al. (2022). Spatially restricted drivers and transitional cell populations cooperate with the microenvironment in untreated and chemo-resistant pancreatic cancer. Nat. Genet. *54*, 1390–1405. https://doi.org/10.1038/s41588-022-01157-1.

90. Travaglini, K.J., Nabhan, A.N., Penland, L., Sinha, R., Gillich, A., Sit, R.V., Chang, S., Conley, S.D., Mori, Y., Seita, J., et al. (2020). A molecular cell atlas of the human lung from single-cell RNA sequencing. Nature *587*, 619–625. https://doi.org/10.1038/s41586-020-2922-4.

91. Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics *25*, 1754–1760. https://doi.org/10.1093/bioinformatics/btp324.

92. Scott, A.D., Huang, K.L., Weerasinghe, A., Mashl, R.J., Gao, Q., Martins Rodrigues, F., Wyczalkowski, M.A., and Ding, L. (2019). CharGer: clinical characterization of germline variants. Bioinformatics *35*, 865–867. https://doi.org/10.1093/bioinformatics/bty649.

93. Newman, A.M., Steen, C.B., Liu, C.L., Gentles, A.J., Chaudhuri, A.A., Scherer, F., Khodadoust, M.S., Esfahani, M.S., Luca, B.A., Steiner, D., et al. (2019). Determining cell type abundance and expression from bulk tissues with digital cytometry. Nat. Biotechnol. *37*, 773–782. https://doi.org/10.1038/s41587-019-0114-2.

94. Taylor-Weiner, A., Stewart, C., Giordano, T., Miller, M., Rosenberg, M., Macbeth, A., Lennon, N., Rheinbay, E., Landau, D.A., Wu, C.J., and Getz, G. (2018). DeTiN: overcoming tumor-in-normal contamination. Nat. Methods *15*, 531–534. https://doi.org/10.1038/s41592-018-0036-9.

95. Mermel, C.H., Schumacher, S.E., Hill, B., Meyerson, M.L., Beroukhim, R., and Getz, G. (2011). GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. Genome Biol. *12*, R41. https://doi.org/10.1186/gb-2011-12-4-r41.

96. Wilkerson, M.D., and Hayes, D.N. (2010). ConsensusClusterPlus: a class discovery tool with confidence assessments and item tracking. Bioinformatics *26*, 1572–1573. https://doi.org/10.1093/bioinformatics/btq170.

97. Alexandrov, L.B., Kim, J., Haradhvala, N.J., Huang, M.N., Tian Ng, A.W., Wu, Y., Boot, A., Covington, K.R., Gordenin, D.A., Bergstrom, E.N., et al. (2020). The repertoire of mutational signatures in human cancer. Nature *578*, 94–101. https://doi.org/10.1038/s41586-020-1943-3.

98. Sun, S.Q., Mashl, R.J., Sengupta, S., Scott, A.D., Wang, W., Batra, P., Wang, L.B., Wyczalkowski, M.A., and Ding, L. (2018). Database of evidence for precision oncology portal. Bioinformatics *34*, 4315–4317. https://doi.org/10.1093/bioinformatics/bty531.

99. Benelli, M., Pescucci, C., Marseglia, G., Severgnini, M., Torricelli, F., and Magi, A. (2012). Discovering chimeric transcripts in paired-end RNA-seq data by using EricScript. Bioinformatics *28*, 3232–3239. https://doi.org/10.1093/bioinformatics/bts617.

100. Anders, S., Pyl, P.T., and Huber, W. (2015). HTSeq–a Python framework to work with high-throughput sequencing data. Bioinformatics *31*, 166–169. https://doi.org/10.1093/bioinformatics/btu638.

101. Zhang, J., White, N.M., Schmidt, H.K., Fulton, R.S., Tomlinson, C., Warren, W.C., Wilson, R.K., and Maher, C.A. (2016). INTEGRATE: gene fusion discovery using whole genome and transcriptome data. Genome Res. *26*, 108–118. https://doi.org/10.1101/gr.186114.114.

102. Chen, X., Schulz-Trieglaff, O., Shaw, R., Barnes, B., Schlesinger, F., Källberg, M., Cox, A.J., Kruglyak, S., and Saunders, C.T. (2016). Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. Bioinformatics *32*, 1220–1222. https://doi.org/10.1093/bioinformatics/btv710.

103. Cibulskis, K., Lawrence, M.S., Carter, S.L., Sivachenko, A., Jaffe, D., Sougnez, C., Gabriel, S., Meyerson, M., Lander, E.S., and Getz, G. (2013). Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. Nat. Biotechnol. *31*, 213–219. https://doi.org/10.1038/nbt.2514.

104. Ye, K., Schulz, M.H., Long, Q., Apweiler, R., and Ning, Z. (2009). Pindel: A pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. Bioinformatics *25*, 2865–2871. https://doi.org/10.1093/bioinformatics/btp394.

105. Polpitiya, A.D., Qian, W.J., Jaitly, N., Petyuk, V.A., Adkins, J.N., Camp, D.G., 2nd, Anderson, G.A., and Smith, R.D. (2008). DAnTE: a statistical tool for quantitative analysis of -omics data. Bioinformatics *24*, 1556–1558. https://doi.org/10.1093/bioinformatics/btn217.

106. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R.; 1000 Genome Project Data Processing Subgroup (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics *25*, 2078–2079. https://doi.org/10.1093/bioinformatics/btp352.

107. Haas, B.J., Dobin, A., Li, B., Stransky, N., Pochet, N., and Regev, A. (2019). Accuracy assessment of fusion transcript detection via read-mapping and de novo fusion transcript assembly-based methods. Genome Biol. *20*, 213. https://doi.org/10.1186/s13059-019-1842-9.

108. Kim, S., Scheffler, K., Halpern, A.L., Bekritsky, M.A., Noh, E., Källberg, M., Chen, X., Kim, Y., Beyter, D., Krusche, P., and Saunders, C.T. (2018). Strelka2: fast and accurate calling of germline and somatic variants. Nat. Methods *15*, 591–594. https://doi.org/10.1038/s41592-018-0051-x.

109. Johnson, J.L., Yaron, T.M., Huntsman, E.M., Kerelsky, A., Song, J., Regev, A., Lin, T.Y., Liberatore, K., Cizin, D.M., Cohen, B.M., et al. (2023). An atlas of substrate specificities for the human serine/threonine kinome. Nature *613*, 759–766. https://doi.org/10.1038/s41586-022-05575-3.

110. Conway, J.R., Lex, A., and Gehlenborg, N. (2017). UpSetR: an R package for the visualization of intersecting sets and their properties. Bioinformatics *33*, 2938–2940. https://doi.org/10.1093/bioinformatics/btx364.

111. Koboldt, D.C., Zhang, Q., Larson, D.E., Shen, D., McLellan, M.D., Lin, L., Miller, C.A., Mardis, E.R., Ding, L., and Wilson, R.K. (2012). VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. Genome Res. *22*, 568–576. https://doi.org/10.1101/gr.129684.111.

112. Aran, D., Hu, Z., and Butte, A.J. (2017). xCell: digitally portraying the tissue cellular heterogeneity landscape. Genome Biol. *18*, 220. https://doi.org/10.1186/s13059-017-1349-1.

113. Haradhvala, N.J., Polak, P., Stojanov, P., Covington, K.R., Shinbrot, E., Hess, J.M., Rheinbay, E., Kim, J., Maruvka, Y.E., Braunstein, L.Z., et al. (2016). Mutational strand asymmetries in cancer genomes reveal mechanisms of DNA damage and repair. Cell *164*, 538–549. https://doi.org/10.1016/j.cell.2015.12.050.

114. McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., and DePristo, M.A. (2010). The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. *20*, 1297–1303. https://doi.org/10.1101/gr.107524.110.

115. McLaren, W., Gil, L., Hunt, S.E., Riat, H.S., Ritchie, G.R., Thormann, A., Flicek, P., and Cunningham, F. (2016). The Ensembl variant effect predictor. Genome Biol. *17*, 122. https://doi.org/10.1186/s13059-016-0974-4.

116. Richards, S., Aziz, N., Bale, S., Bick, D., Das, S., Gastier-Foster, J., Grody, W.W., Hegde, M., Lyon, E., Spector, E., et al. (2015). Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. Genet. Med. *17*, 405–424. https://doi.org/10.1038/gim.2015.30.

117. Kumar, P., Henikoff, S., and Ng, P.C. (2009). Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. Nat. Protoc. *4*, 1073–1081. https://doi.org/10.1038/nprot.2009.86.

118. Adzhubei, I., Jordan, D.M., and Sunyaev, S.R. (2013). Chapter 7. Predicting functional effect of human missense mutations using PolyPhen-2Unit7.20. Curr. Protoc. Hum. Genet. *Chapter 7*. https://doi.org/10.1002/0471142905.hg0720s76.

119. Huang, K.L., Mashl, R.J., Wu, Y., Ritter, D.I., Wang, J., Oh, C., Paczkowska, M., Reynolds, S., Wyczalkowski, M.A., Oak, N., et al. (2018). Pathogenic germline variants in 10,389 adult cancers. Cell *173*, 355–370.e14. https://doi.org/10.1016/j.cell.2018.03.039.

120. Karczewski, K.J., Francioli, L.C., Tiao, G., Cummings, B.B., Alföldi, J., Wang, Q., Collins, R.L., Laricchia, K.M., Ganna, A., Birnbaum, D.P., et al. (2020). The mutational constraint spectrum quantified from variation in 141,456 humans. Nature *581*, 434–443. https://doi.org/10.1038/s41586-020-2308-7.

121. 1000 Genomes Project Consortium, Auton, A., Brooks, L.D., Durbin, R.M., Garrison, E.P., Kang, H.M., Korbel, J.O., Marchini, J.L., McCarthy, S., McVean, G.A., et al. (2015). A global reference for human genetic variation. Nature *526*, 68–74. https://doi.org/10.1038/nature15393.

122. Xi, R., Lee, S., Xia, Y., Kim, T.M., and Park, P.J. (2016). Copy number analysis of whole-genome data using BIC-seq2 and its application to detection of cancer susceptibility variants. Nucleic Acids Res. *44*, 6274–6286. https://doi.org/10.1093/nar/gkw491.

123. Saghafinia, S., Mina, M., Riggi, N., Hanahan, D., and Ciriello, G. (2018). Pan-cancer landscape of aberrant DNA methylation across human tumors. Cell Rep. *25*, 1066–1080.e8. https://doi.org/10.1016/j.celrep.2018.09.082.

124. Gao, Q., Liang, W.W., Foltz, S.M., Mutharasu, G., Jayasinghe, R.G., Cao, S., Liao, W.W., Reynolds, S.M., Wyczalkowski, M.A., Yao, L., et al. (2018). Driver fusions and their implications in the development and treatment of human cancers. Cell Rep. *23*, 227–238.e3. https://doi.org/10.1016/j.celrep.2018.03.050.

125. Babiceanu, M., Qin, F., Xie, Z., Jia, Y., Lopez, K., Janus, N., Facemire, L., Kumar, S., Pang, Y., Qi, Y., et al. (2016). Recurrent chimeric fusion RNAs in non-cancer tissues and cells. Nucleic Acids Res. *44*, 2859–2872. https://doi.org/10.1093/nar/gkw032.

126. Hao, Y., Hao, S., Andersen-Nissen, E., Mauck, W.M., 3rd, Zheng, S., Butler, A., Lee, M.J., Wilk, A.J., Darby, C., Zager, M., et al. (2021). Integrated analysis of multimodal single-cell data. Cell *184*, 3573–3587.e29. https://doi.org/10.1016/j.cell.2021.04.048.

127. Yoshihara, K., Shahmoradgoli, M., Martínez, E., Vegesna, R., Kim, H., Torres-Garcia, W., Treviño, V., Shen, H., Laird, P.W., Levine, D.A., et al. (2013). Inferring tumour purity and stromal and immune cell admixture from expression data. Nat. Commun. *4*, 2612. https://doi.org/10.1038/ncomms3612.

128. 1000 Genomes Project Consortium, Abecasis, G.R., Altshuler, D., Auton, A., Brooks, L.D., Durbin, R.M., Gibbs, R.A., Hurles, M.E., and McVean, G.A. (2010). A map of human genome variation from population-scale sequencing. Nature *467*, 1061–1073. https://doi.org/10.1038/nature09534.

129. Andreatta, M., and Nielsen, M. (2016). Gapped sequence alignment using artificial neural networks: application to the MHC class I system. Bioinformatics *32*, 511–517. https://doi.org/10.1093/bioinformatics/btv639.

130. Kundra, R., Zhang, H., Sheridan, R., Sirintrapun, S.J., Wang, A., Ochoa, A., Wilson, M., Gross, B., Sun, Y., Madupuri, R., et al. (2021). OncoTree: A cancer classification system for precision oncology. JCO Clin. Cancer Inform. *5*, 221–230. https://doi.org/10.1200/CCI.20.00108.

131. Sondka, Z., Bamford, S., Cole, C.G., Ward, S.A., Dunham, I., and Forbes, S.A. (2018). The COSMIC Cancer Gene Census: describing genetic dysfunction across all human cancers. Nat. Rev. Cancer *18*, 696–705. https://doi.org/10.1038/s41568-018-0060-1.

132. Ellrott, K., Bailey, M.H., Saksena, G., Covington, K.R., Kandoth, C., Stewart, C., Hess, J., Ma, S., Chiotti, K.E., McLellan, M., et al. (2018). Scalable open science approach for mutation calling of tumor exomes using multiple genomic pipelines. Cell Syst. *6*, 271–281.e7. https://doi.org/10.1016/j.cels.2018.03.002.

133. Pagel, K.A., Kim, R., Moad, K., Busby, B., Zheng, L., Tokheim, C., Ryan, M., and Karchin, R. (2020). Integrated informatics analysis of cancer-related variants. JCO Clin. Cancer Inform. *4*, 310–317. https://doi.org/10.1200/CCI.19.00132.

134. Ng, P.C., and Henikoff, S. (2002). Accounting for human polymorphisms predicted to affect protein function. Genome Res. *12*, 436–446. https://doi.org/10.1101/gr.212802.

135. Carter, H., Douville, C., Stenson, P.D., Cooper, D.N., and Karchin, R. (2013). Identifying Mendelian disease genes with the variant effect scoring tool. BMC Genomics *14 Suppl 3*, S3. https://doi.org/10.1186/1471-2164-14-S3-S3.

136. Reva, B., Antipin, Y., and Sander, C. (2011). Predicting the functional impact of protein mutations: application to cancer genomics. Nucleic Acids Res. *39*, e118. https://doi.org/10.1093/nar/gkr407.

137. Mao, Y., Chen, H., Liang, H., Meric-Bernstam, F., Mills, G.B., and Chen, K. (2013). CanDrA: cancer-specific driver missense mutation annotation with optimized features. PLoS One *8*, e77945. https://doi.org/10.1371/journal.pone.0077945.

138. Shihab, H.A., Gough, J., Cooper, D.N., Stenson, P.D., Barker, G.L., Edwards, K.J., Day, I.N., and Gaunt, T.R. (2013). Predicting the functional, molecular, and phenotypic consequences of amino acid substitutions using hidden Markov models. Hum. Mutat. *34*, 57–65. https://doi.org/10.1002/humu.22225.

139. Carter, H., Chen, S., Isik, L., Tyekucheva, S., Velculescu, V.E., Kinzler, K.W., Vogelstein, B., and Karchin, R. (2009). Cancer-specific high-throughput annotation of somatic mutations: computational prediction of driver missense mutations. Cancer Res. *69*, 6660–6667. https://doi.org/10.1158/0008-5472.CAN-09-1133.

140. Gonzalez-Perez, A., Deu-Pons, J., and Lopez-Bigas, N. (2012). Improving the prediction of the functional impact of cancer mutations by baseline tolerance transformation. Genome Med. *4*, 89. https://doi.org/10.1186/gm390.

141. Niu, B., Scott, A.D., Sengupta, S., Bailey, M.H., Batra, P., Ning, J., Wyczalkowski, M.A., Liang, W.W., Zhang, Q., McLellan, M.D., et al. (2016). Protein-structure-guided discovery of functional mutations across 19 cancer types. Nat. Genet. *48*, 827–837. https://doi.org/10.1038/ng.3586.

142. Gao, J., Chang, M.T., Johnsen, H.C., Gao, S.P., Sylvester, B.E., Sumer, S.O., Zhang, H., Solit, D.B., Taylor, B.S., Schultz, N., and Sander, C. (2017). 3D clusters of somatic mutations in cancer reveal numerous rare mutations as functional targets. Genome Med. *9*, 4. https://doi.org/10.1186/s13073-016-0393-x.

143. Chang, M.T., Bhattarai, T.S., Schram, A.M., Bielski, C.M., Donoghue, M.T.A., Jonsson, P., Chakravarty, D., Phillips, S., Kandoth, C., Penson, A., et al. (2018). Accelerating discovery of functional mutant alleles in cancer. Cancer Discov. *8*, 174–183. https://doi.org/10.1158/2159-8290.CD-17-0321.

144. Vogelstein, B., Papadopoulos, N., Velculescu, V.E., Zhou, S., Diaz, L.A., Jr., and Kinzler, K.W. (2013). Cancer genome landscapes. Science *339*, 1546–1558. https://doi.org/10.1126/science.1235122.

145. Tokheim, C.J., Papadopoulos, N., Kinzler, K.W., Vogelstein, B., and Karchin, R. (2016). Evaluating the evaluation of cancer driver genes. Proc. Natl. Acad. Sci. USA. *113*, 14330–14335. https://doi.org/10.1073/pnas.1616440113.

146. Murphy, C., and Elemento, O. (2016). AGFusion: annotate and visualize gene fusions. bioRxiv.. https://doi.org/10.1101/080903.

147. Varadi, M., Anyango, S., Deshpande, M., Nair, S., Natassia, C., Yordanova, G., Yuan, D., Stroe, O., Wood, G., Laydon, A., et al. (2022). AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. Nucleic Acids Res. *50*, D439–D444. https://doi.org/10.1093/nar/gkab1061.

148. Kabsch, W., and Sander, C. (1983). Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers *22*, 2577–2637. https://doi.org/10.1002/bip.360221211.

149. Rost, B., and Sander, C. (1994). Conservation and prediction of solvent accessibility in protein families. Proteins *20*, 216–226. https://doi.org/10.1002/prot.340200303.

150. Szklarczyk, D., Gable, A.L., Nastou, K.C., Lyon, D., Kirsch, R., Pyysalo, S., Doncheva, N.T., Legeay, M., Fang, T., Bork, P., et al. (2021). The STRING database in 2021: customizable protein-protein networks, and functional characterization of user-uploaded gene/measurement sets. Nucleic Acids Res. *49*, D605–D612. https://doi.org/10.1093/nar/gkaa1074.

151. Calderone, A., Iannuccelli, M., Peluso, D., and Licata, L. (2020). Using the MINT database to search protein interactions. Curr. Protoc. Bioinformatics *69*, e93. https://doi.org/10.1002/cpbi.93.

152. Oughtred, R., Stark, C., Breitkreutz, B.J., Rust, J., Boucher, L., Chang, C., Kolas, N., O'Donnell, L., Leung, G., McAdam, R., et al. (2019). The BioGRID interaction database: 2019 update. Nucleic Acids Res. *47*, D529–D541. https://doi.org/10.1093/nar/gky1079.

153. Del Toro, N., Shrivastava, A., Ragueneau, E., Meldal, B., Combe, C., Barrera, E., Perfetto, L., How, K., Ratan, P., Shirodkar, G., et al. (2022). The IntAct database: efficient access to fine-grained molecular interaction data. Nucleic Acids Res. *50*, D648–D653. https://doi.org/10.1093/nar/gkab1006.

154. Giurgiu, M., Reinhard, J., Brauner, B., Dunger-Kaltenbach, I., Fobo, G., Frishman, G., Montrone, C., and Ruepp, A. (2019). CORUM: The comprehensive resource of mammalian protein complexes-2019. Nucleic Acids Res. *47*, D559–D563. https://doi.org/10.1093/nar/gky973.

155. Luck, K., Kim, D.K., Lambourne, L., Spirohn, K., Begg, B.E., Bian, W., Brignall, R., Cafarelli, T., Campos-Laborie, F.J., Charloteaux, B., et al. (2020). A reference map of the human binary protein interactome. Nature *580*, 402–408. https://doi.org/10.1038/s41586-020-2188-x.

156. Flanders, W.D., DerSimonian, R., and Freedman, D.S. (1992). Interpretation of linear regression models that include transformations or interaction terms. Ann. Epidemiol. *2*, 735–744. https://doi.org/10.1016/1047-2797(92)90018-L.

157. Fritz, M.S., and Mackinnon, D.P. (2007). Required sample size to detect the mediated effect. Psychol. Sci. *18*, 233–239. https://doi.org/10.1111/j.1467-9280.2007.01882.x.

158. Steen, J., Loeys, T., Moerkerke, B., and Vansteelandt, S. (2017). medflex: an R package for flexible mediation analysis using natural effect models. J. Stat. Soft. *76*, 1–46. https://doi.org/10.18637/jss.v076.i11.

159. Hornbeck, P.V., Zhang, B., Murray, B., Kornhauser, J.M., Latham, V., and Skrzypek, E. (2015). PhosphoSitePlus, 2014: mutations, PTMs and recalibrations. Nucleic Acids Res. *43*, D512–D520. https://doi.org/10.1093/nar/gku1267.

160. Chen, B., Khodadoust, M.S., Liu, C.L., Newman, A.M., and Alizadeh, A.A. (2018). Profiling tumor infiltrating immune cells with CIBERSORT. Methods Mol. Biol. *1711*, 243–259. https://doi.org/10.1007/978-1-4939-7493-1_12.

161. Jiang, P., Gu, S., Pan, D., Fu, J., Sahu, A., Hu, X., Li, Z., Traugh, N., Bu, X., Li, B., et al. (2018). Signatures of T cell dysfunction and exclusion predict cancer immunotherapy response. Nat. Med. *24*, 1550–1558. https://doi.org/10.1038/s41591-018-0136-1.

162. ENCODE Project Consortium (2012). An integrated encyclopedia of DNA elements in the human genome. Nature *489*, 57–74. https://doi.org/10.1038/nature11247.

163. Chicas, A., Wang, X., Zhang, C., McCurrach, M., Zhao, Z., Mert, O., Dickins, R.A., Narita, M., Zhang, M., and Lowe, S.W. (2010). Dissecting the unique role of the retinoblastoma tumor suppressor during cellular senescence. Cancer Cell *17*, 376–387. https://doi.org/10.1016/j.ccr.2010.01.023.

164. Ferrari, R., Gou, D., Jawdekar, G., Johnson, S.A., Nava, M., Su, T., Yousef, A.F., Zemke, N.R., Pellegrini, M., Kurdistani, S.K., and Berk, A.J. (2014). Adenovirus small E1A employs the lysine acetylases p300/CBP and tumor suppressor Rb to repress select host genes and promote productive virus infection. Cell Host Microbe *16*, 663–676. https://doi.org/10.1016/j.chom.2014.10.004.

165. Zheng, R., Wan, C., Mei, S., Qin, Q., Wu, Q., Sun, H., Chen, C.H., Brown, M., Zhang, X., Meyer, C.A., and Liu, X.S. (2019). Cistrome data browser: expanded datasets and new tools for gene regulatory analysis. Nucleic Acids Res. *47*, D729–D735. https://doi.org/10.1093/nar/gky1094.

166. Wang, S., Zang, C., Xiao, T., Fan, J., Mei, S., Qin, Q., Wu, Q., Li, X., Xu, K., He, H.H., et al. (2016). Modeling cis-regulation with a compendium of genome-wide histone H3K27ac profiles. Genome Res. *26*, 1417–1429. https://doi.org/10.1101/gr.201574.115.

167. Stoney, R.A., Schwartz, J.M., Robertson, D.L., and Nenadic, G. (2018). Using set theory to reduce redundancy in pathway sets. BMC Bioinformatics *19*, 386. https://doi.org/10.1186/s12859-018-2355-3.

168. Geffen, Y., Anand, S., Akiyama, Y., Yaron, T.M., Song, Y., Johnson, J.L., Govindan, A., Babur, O., Li, Y., Huntsman, E., et al. (2023). Pan-cancer analysis of post-translational modifications reveals shared patterns of protein regulation. Cell *186*, 3945–3967.

169. Lawrence, M.S., Stojanov, P., Mermel, C.H., Robinson, J.T., Garraway, L.A., Golub, T.R., Meyerson, M., Gabriel, S.B., Lander, E.S., and Getz, G. (2014). Discovery and saturation analysis of cancer genes across 21 tumour types. Nature *505*, 495–501. https://doi.org/10.1038/nature12912.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
CPTAC clinical data and proteomic data	Li et al.[87]	Proteomic Data Commons (PDC: https://pdc.cancer.gov/pdc/cptac-pancancer)
CPTAC genomic, transcriptomic data	Li et al.[87]	Proteomic Data Commons (PDC:https://pdc.cancer.gov/pdc/cptac-pancancer), and Cancer Data Service (CDS: https://dataservice.datacommons.cancer.gov/)
CPTAC pathology and radiology images	Li et al.[87]	Imaging Data Commons (https://portal.imaging.datacommons.cancer.gov/explore/)
Full proteomic data tables	This manuscript	Proteomic Data Commons (PDC: https://pdc.cancer.gov/pdc/cptac-pancancer)
TCGA	Bailey et al.[1]	Genomic Data Commons (GDC:)
single nuclei RNA sequencing and single-cell RNA sequencing	Wang et al.,[16] Li et al.,[88] Cui Zhou et al.,[89] and Travaglini et al.[90]	https://portal.gdc.cancer.gov/projects/CPTAC-3 https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs002371.v1.p1 https://www.synapse.org/#!Synapse:syn21041850
Software and algorithms		
3Dmapper	Porta-Pardo Lab	https://github.com/vicruiser/3Dmapper
Ancestry prediction	Li Ding Lab	https://github.com/ding-lab/ancestry
BWA v0.7.17-r1188	Li and Durbin[91]	http://bio-bwa.sourceforge.net/
C3PO	This manuscript	https://github.com/MHBailey/C3PO_polygenic
CharGer v.0.5.4	Scott et al.[92]	https://github.com/ding-lab/CharGer/tree/v0.5.4
CIBERSORTx	Newman et al.[93]	https://cibersortx.stanford.edu/
WES hg38 characterization pipeline	Getz Lab	https://terra.bio/
bam-readcount v0.8	McDonnell Genome Institute	https://github.com/genome/bam-readcount
GATK4's CalculateContamination	GATK	https://gatk.broadinstitute.org/hc/en-us/articles/360036888972-CalculateContamination
GATK4 Picard tools	GATK	https://github.com/broadinstitute/picard
GATK4 Funcotator	GATK	https://gatk.broadinstitute.org/hc/en-us/articles/360037224432-Funcotator
DeTiN	Taylor-Weiner et al.[94]	https://github.com/getzlab/deTiN
Spectrum Mill	Karl R. Clauser, Steven Carr Lab	https://proteomics.broadinstitute.org/
GISTIC2.0	Mermel et al.[95]	https://github.com/broadinstitute/gistic2
ConsensusClusterPlus v1.48.0	Wilkerson and Hayes[96]	https://bioconductor.org/packages/ConsensusClusterPlus/
COSMIC Mutational Signatures v3	Alexandrov et al.[97]	https://cancer.sanger.ac.uk/cosmic/signatures/
DEPO	Sun et al.[98]	http://dinglab.wustl.edu/depo
EricScript v0.5.5	Benelli et al.[99]	https://sites.google.com/site/bioericscript/
germlinewrapper v1.1	Li Ding Lab	https://github.com/ding-lab/germlinewrapper
HTSeq v0.11.2	Anders et al.[100]	https://github.com/simon-anders/htseq
INTEGRATE v0.2.6	Zhang et al.[101]	https://sourceforge.net/projects/integrate-fusion/
Manta v1.6.0	Chen et al.[102]	https://github.com/Illumina/manta
Methylation analysis	Li Ding Lab	https://github.com/ding-lab/cptac_methylation
MSI prediction	Li Ding Lab	https://github.com/ding-lab/msisensor
MuTect v1.1.7	Cibulskis et al.[103]	https://github.com/broadinstitute/mutect
Pindel v0.2.5	Ye et al.[104]	https://github.com/genome/pindel

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Python v3.7	Python Software Foundation	https://www.python.org/
R v3.6	R Development Core Team	https://www.R-project.org
R-rollup	Polpitiya et al.[105]	https://omics.pnl.gov/software/danter
Samtools v1.2	Li et al.[106]	https://www.htslib.org/
SignatureAnalyzer	Alexandrov et al.[97]	https://github.com/broadinstitute/getzlab-SignatureAnalyzer
somaticwrapper v1.6	Li Ding Lab	https://github.com/ding-lab/somaticwrapper
STAR-Fusion v1.5.0	Haas et al.[107]	https://github.com/STAR-Fusion/STAR-Fusion
Strelka v2.9.2	Kim et al.[108]	https://github.com/Illumina/strelka
the Kinase Library 2.4.0	Johnson et al.[109]	https://kinase-library.phosphosite.org/
UpSetR	Conway et al.[110]	https://github.com/hms-dbmi/UpSetR/
VarScan v2.3.8	Koboldt et al.[111]	https://dkoboldt.github.io/varscan/
xCell v1.2	Aran et al.[112]	http://xcell.ucsf.edu/

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the Lead Contact, Li Ding (lding@wustl.edu).

Materials availability

This study did not generate new unique reagents.

Data and code availability

- Raw and processed proteomics as well as open-access genomic data, can be obtained via Proteomic Data Commons (PDC) at https://pdc.cancer.gov/pdc/cptac-pancancer. Raw genomic and transcriptomic data files can be accessed via the Genomic Data Commons (GDC) Data Portal at https://portal.gdc.cancer.gov with dbGaP Study Accession: phs001287.v16.p6. Complete CPTAC pan-cancer controlled and processed data can be accessed via the Cancer Data Service (CDS: https://dataservice.datacommons.cancer.gov/). The CPTAC pan-cancer data hosted in CDS is controlled data and can be accessed through the NCI DAC approved, dbGaP compiled whitelists. Users can access the data for analysis through the Seven Bridges Cancer Genomics Cloud (SB-CGC) which is one of the NCI-funded Cloud Resource/platform for compute intensive analysis. Instructions to access data: 1. Create an account on CGC, Seven Bridges (https://cgc-accounts.sbgenomics.com/auth/register 2. Get approval from dbGaP to access the controlled study (https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs001287.v16.p6) 3. Log into CGC to access Cancer Data Service (CDS) File Explore 4. Copy data into your own space and start analysis and exploration 5. Visit the CDS page on CGC to see what studies are available and instructions and guides to use the resources. (https://docs.cancergenomicscloud.org/page/cds-data). We focused on the CPTAC samples with both genomic and proteomic data available to investigate the pan-cancer proteogenomic impacts of oncogenic drivers.
- All original code has been deposited at GitHub and is publicly available as of the date of publication. DOIs are listed in the key resources table.
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Human subjects

A total of 1064 participants were included in this study. Prospective biospecimen collection (tumor, germline blood and adjacent normal samples where feasible) followed a tumor type specific protocol and standard operating procedures (SOPs), where sample collection, qualification and processing were optimized for both genomics and proteomics.[13–22] CPTAC samples were collected by 30+ tissue source sites from both domestic and international locations and processed by a central biospecimen core resource. The samples were pathology qualified by a general pathologist and later reconfirmed by a disease-specific expert pathologist through histopathology image review and immunohistochemistry assays where applicable.

Clinical data annotation

Clinical data were obtained from TSS and aggregated by the Biospecimen Core Resource (BCR, Van Andel Research Institute (Grand Rapids, MI)). Data forms were stored as Microsoft Excel files (.xls). Clinical data can be accessed and downloaded from the CPTAC Data Portal and https://pdc.cancer.gov/pdc/cptac-pancancer as described in Li et al.[87].

METHOD DETAILS

Harmonized genome alignment

WGS, WES, and RNA-Seq sequence data were harmonized by NCI Genomic Data Commons (GDC) https://gdc.cancer.gov/about-data/gdc-data-harmonization, which included alignment to GDC's hg38 human reference genome (GRCh38.d1.vd1) and additional quality checks. All the downstream genomic processing was based on the GDC-aligned BAMs to ensure reproducibility.

Somatic mutation calling

Somatic Mutation Calling (pipeline from Washington University in St Louis)

Somatic mutations were called by the Somaticwrapper pipeline v1.6 (https://github.com/ding-lab/somaticwrapper), which includes four different callers, i.e., Strelka v.2,[108] MUTECT v1.1.7,[103] VarScan v.2.3.8,[111] and Pindel v.0.2.5[104] from WES. We kept the exonic SNVs called by any two callers among MUTECT v1.7, VarScan v.2.3.8, and Strelka v.2.9.2 and indels called by any two callers among VarScan v.2.3.8, Strelka v.2, and Pindel v.0.2.5. For the merged SNVs and indels, we applied a 14X and 8X coverage cutoff for tumor and normal, separately. We also filtered SNVs and indels by a minimal variant allele frequency (VAF) of 0.05 in tumors and a maximal VAF of 0.02 in normal samples. We filtered any SNV within 10bp of an indel found in the same tumor sample. Finally, we rescued the rare mutations with VAF of [0.015, 0.05) in ccRCC driver genes based on the gene consensus list.[1]

Somatic mutation calling (pipeline from Broad Institute of MIT and Harvard)

In parallel, patient whole-exome sequencing (WES) data for matched tumor/normal samples were analyzed using the Getz Lab's production hg38 WES characterization pipeline. The hg38 characterization pipeline runs on the Terra cloud-based analysis platform (https://terra.bio/). This pipeline is the Getz Lab's standard computational workflow, and the analysis steps are organized into five modules: (1) DNA Sequence Data Quality Control (including GATK4's CalculateContamination [ver GATK 4.1.4.1] and GATK4 Picard tools [ver GATK 4.0.5.1]). (2) Somatic Copy Number Analysis (GATK4 Best Practices Workflow [ver GATK 4.1.4.1]). (3) Somatic Variant Discovery, which includes the discovery of single-nucleotide variants (SNVs) and insertions/deletions (indels), using MuTect[103] and Manta+Strelka v2.[102,108] Next, deTiN v1.8.9[94] is run to account for and rescue tumor-in-normal contamination. The resulting SNV and indel VCFs are each run through the GATK4 Funcotator (ver GATK 4.1.4.1). (4) Post-Discovery Filtering, which employs a collection of filters to remove alignment artifacts, germline variants, and common sequencing artifacts that occur in normal panels. (5) Merging of adjacent somatic Single Nucleotide Polymorphisms (SNPs) into di-nucleotide polymorphisms (DNPs), trinucleotide polymorphisms (TNPs), and Oligo-nucleotide polymorphisms (ONPs).

Somatic mutation callset harmonization (pipeline from Broad Institute of MIT and Harvard)

The per-patient variant calls employed by the CPTAC PanCan working group are derived from the harmonization of variant calls made independently by the Broad and Washthe U teams. First, we filter calls outside of the inosine chemical erasing (ICE) interval (Genomics Platform at the Broad Institute) and apply a "panel-of-normals" built from aggregating CPTAC and TCGA cohorts for consistency between the two pipelines. In addition, to account for differences in the two mutation call sets, we: (i) removed all calls with Variant Allele Frequency (VAF)<0.05 from both pipelines and rescued only high confident calls and (ii) long ONPs were collapsed to shorter ONPs by imposing a more stringent merging criterion that requires a 2bp gap length at max. Next, we used Asymtools2[113] to identify a sequencing artifact affecting CPTAC2 whole-exome sequencing. We then corrected for the sequencing artifact by ranking context-specific mutations by their allelic fractions. Finally, the functional impact of harmonized calls was annotated using GATK Funcotator.

Germline variant calling and pathogenicity classification

Germline variant calling was performed using the GermlineWrapper v1.1 pipeline, which implements multiple tools for detecting germline INDELs and SNVs. Germline SNVs were identified using VarScan v2.3.8[111] (with parameters:-min-var-freq 0.10, -p-value 0.10, -min-coverage 3, -strand-filter 1) operating on a mpileup stream produced by samtools v1.2 (with parameters: -q 1 -Q 13) and GATK v4.0.0.0[114] using its haplotype caller in single-sample mode with duplicate and unmapped reads removed and retaining calls with a minimum quality threshold of 10. Germline INDELs were called using VarScan, GATK, and Pindel[104] (version 0.2.5b9, with default parameters except -m 6, -w 1, excluding centromere regions (genome.ucsc.edu)). INDELs were required to be called by at least two out of the three callers. All resulting variants were limited to coding regions of full-length transcripts obtained from Ensembl release 100 plus additional two base pairs flanking each exon to cover splice sites. We required variants to have Allelic Depth (AD) $\geq$ 5 reads for the alternative allele and filtered out any INDELs longer than 100bps. Germline variants passing filters were then annotated using the Ensembl Variant effect Predictor (VEP)[115] (version 100 with default parameters, except -everything) and their pathogenicity was determined with our automated pipeline CharGer[92] (version 0.5.4), which classifies germline variants based on guidelines published by the American College of Medical Genetics (ACMG).[116] Briefly, CharGer applies 12 pathogenic and four benign evidence levels using information from several databases, such as gnomAD and ClinVAr, and *in silico* tools, such as SIFT[117] and Polyphen.[118] Further details on implementation, scores of each evidence level, and parameters used are the same as those in our previous

pan-cancer TCGA germline study.[92,119] Variants classified by CharGer were then filtered for rare variants with minor allele frequency (MAF) $\leq$ 0.05% in gnomAD r.2.1.1[120] and/or The 1000 Genomes Project.[121] Further, we used bam-readcount v0.8 for reference and alternative alleles quantification (with parameters: -q 10 -b 15) in both normal and tumor samples and required at least five counts for the alternative allele and 20% variant allele frequency (VAF) in both tumor and normal samples. Variants passing these filters were = manually reviewed with the Integrative Genomics Viewer (IGV) software (version 2.12.3) using = tumor and normal data.

Copy number variant calling

Genomic data post-processing and GISTIC (pipeline from Broad Institute of MIT and Harvard)

The Genomic Identification of Significant Targets in Cancer (GISTIC2.0) algorithm[95] was used to identify significantly amplified or deleted focal-level and arm-level events, with q-value<0.25 considered significant (default parameters were used).

WGS copy number variant calling (pipeline from Washington University in St Louis)

We used BIC-seq2,[122] a read-depth-based CNA calling algorithm, to detect somatic copy number alteration (CNAs) from the WGS data of tumors. Briefly, BIC-seq2 divides genomic regions into disjoint bins and counts uniquely aligned reads in each bin. Then, it combines neighboring bins into genomic segments with similar copy numbers iteratively based on Bay'sian Information Criteria (BIC), a statistical criterion measuring a statistical model's fitness and complexity. We used paired-sample CNA calling that takes samples as input and detects genomic regions with different copy numbers between the two samples. We used a bin size of ~100 bp and a lambda of 3 (a smoothing parameter for CNA segmentation). We recommend calling segments as copy, gain, or loss when their $\log_2$ copy ratios are larger than 0.2 or smaller than -0.2, respectively (according to the BIC-seq publication).

DNA methylation microarray processing

Raw methylation idat files were downloaded from CPTAC DCC and GDC. Beta values of CpG loci were reported after functional normalization, quality check, common SNP filtering, and probe annotation using the Li Ding Lab's methylation pipeline v1.1 https://github.com/ding-lab/cptac_methylation.

Gene-level DNA methylation data were generated using the beta values of all probes harboring in the islands of promoter and 5 UTR regions of the genes. Samples with a missing rate>10% across all probes were excluded, the mean and median levels were calculated, and the analyses were performed separately on the discovery cohort of seven cancer types. For five cancer types (ccRCC, LUAD, LSCC, HNSCC, and PDAC) that have DNA methylation data in both NAT and tumor samples, consensus clustering analysis was performed as a quality control procedure to evaluate the separation of clusters on NAT vs. tumor samples. The misclassified samples were checked for clustering uncertainty and sample labeling. We identified aberrant DNA methylation associated with transcriptional and/or translational changes using the published RESET pipeline.[123] Within the promoter region (TSS +- 300 bp).

RNA data processing and quantification

We obtained the gene-level read count, Fragments Per Kilobase of transcript per Million mapped reads (FPKM), and FPKM Upper Quartile (FPKM-UQ) values by following the GDC's RNA-Seq pipeline (Expression mRNA Pipeline) https://docs.gdc.cancer.gov/Data/Bioinformatics_Pipelines/Expression_mRNA_Pipeline/, except running the quantification tools in the stranded mode. We used HTSeq v0.11.2[100] to calculate the gene-level stranded read count (parameters: -r pos -f bam -a 10 -s reverse -t exon -i gene_id -m intersection-nonempty –nonunique=none) using GENCODE v22 (Ensembl v79) annotation downloaded from GDC (gencode.gene.info.v22.tsv). The read count was then converted to FPKM and FPKM-UQ using the formula described in GDC's Expression mRNA Pipeline documentation.

RNA fusion detection

We used three callers, STAR-Fusion v1.5.0,[107] INTEGRATE v0.2.6,[101] and EricScript v0.5.5,[99] to call consensus fusion/chimeric events in our samples. Calls by each tool using tumor and normal RNA-Seq data were then merged into a single file, and extensive filtering was done. As STAR-Fusion has higher sensitivity, calls made by this tool with higher supporting evidence (defined by fusion fragments per million total reads, or FFPM>0.1) were required, or a given fusion must be reported by at least two callers. We then removed the fusions present in our panel of blacklisted or normal fusions, which included uncharacterized genes, immunoglobulin genes, mitochondrial genes, and others, as well as fusions from the same gene or paralog genes and fusions reported in TCGA normal samples,[124] GTEx tissues (reported in STAR-Fusion output), and non-cancer cell studies.[125] Finally, we removed normal fusions from the tumor fusions to curate the final set.

Cell-type annotation on differentially expressed markers

We applied a comprehensive analysis strategy to annotate the cell-type expression of differentially expressed markers. As the primary approach, we utilized the published expression data at the single-cell resolution[16,88–90] and separated the populations into three main categories (e.g., Tumor, Immune, and Stromal). Differentially expressed genes corresponding to each category were identified by the FindMarkers function in Seurat.[126] Wilcoxon rank-sum test was used. $\log_2$FC>1 and q-value<0.05 were used to filter DEGs and classify the differentially expressed markers into Tumor (T), Immune (I), Stromal (S), or Mixed (M) categories. For those that were not quantified in the data described above or did not pass through the cutoffs, we applied secondary annotation by performing the correlation analysis between the marker expression and cell-type-signature (e.g., ESTIMATE[127] and xCell[112]) and applied

the cutoffs of correlation difference>0.5 and q-value<0.05. For example, if a marker has significant correlations (e.g., q-value<0.05) with tumor purity of 0.7, with an immune score of -0.3, and with a stromal score of -0.05, it will be considered as the category of (T). Moreover, for the small subset that was not evaluable by the primary and secondary annotation, we referred to the Human Protein Atlas for their expression profiles as a part of the multi-layer annotations.

Tumor microenvironment inference

The ESTIMATE scores reflecting the overall immune and stromal infiltration were calculated by the R package ESTIMATE[127] using the normalized RNA expression data (FPKM-UQ).

Cell type enrichment deconvolution using gene expression

The abundance of each cell type was inferred by the xCell web tool,[112] which performed the cell type enrichment analysis from gene expression data for 64 immune and stromal cell types (default xCell signature). xCell is a gene signatures-based method learned from thousands of pure cell types from various sources. We used the FPKM-UQ expression matrix as the input of xCell. xCell generated an immune score per sample that integrates the enrichment scores of B cells, CD4+ T-cells, CD8+ T-cells, DC, eosinophils, macrophages, monocytes, mast cells, neutrophils, and NK cells; a micro-environment score which was the sum of the immune score and stroma score. Besides, we applied CIBERSORTx[93] to compute immune cell fractions from bulk gene expression data.

Survival analysis

The R package ''survival'' was used to perform survival analysis. The Kaplan-Meier curve of overall survival was used to compare the prognosis among subtypes (function survfit). Log-rank test (from the R package survminer) was used to test the differential survival outcomes between categorical variables. The standard multivariate Cox-proportional hazard modeling was applied to estimate the hazard ratio among subtypes (function coxph).

Ancestry prediction using SNPs from 1000 genomes project

We used a reference panel of genotypes and a clustering based on principal components to identify likely ancestry. We selected 107,765 coding SNPs with a minor allele frequency>0.02 from the final phase release of The 1000 Genomes Project.[128] From this set of loci, we measured the depth and allele counts of each sample in our cohort using bam-readcount v0.8.0. Genotypes were then called for each sample based on the following criteria: 0/0 if reference count $\geq$ 8 and alternate count<4; 0/1 if reference count $\geq$ 4 and alternate count $\geq$ 4; 1/1 if reference count<4 and alternate count $\geq$ 8; and./. (missing) otherwise. After excluding markers with missingness>5%, 70,968 markers were kept for analysis. We performed PCA on the 1000 Genomes samples to identify the top 20 principal components. We then projected our cohort onto the 20-dimensional space representing the 1000 Genomes data. We then trained a random forest classifier with the 1000 Genomes dataset using these 20 principal components. The 1000 Genomes dataset was split 80/20 for training and validation, respectively. On the validation dataset, our classifier achieved 99.6% accuracy. We then used the fitted classifier to predict the likely ancestry of our cohort. Each prediction file contains classification probabilities for each sample for five ancestries (EUS - European, EAS - East Asian, SAS - South Asian, AFR - African, AMR - American Admixture) GitHub: https://github.com/ding-lab/ancestry.

MSI prediction

MSI scores were calculated by MSIsensor (https://github.com/ding-lab/msisensor) and interpreted as the percentage of microsatellite sites (with deep enough sequencing coverage) that have a lesion. Samples with an MSI score>3.5 are classified as "MSI-high" and the rest will be classified as "MSS." An intermediate class with 1.0<=score<=3.5 can be defined as "MSI-low."

HLA typing and neoantigen prediction

The wild-type/variant protein sequences are obtained from the Refseq database. We constructed different epitope lengths (8-11-mer) from the translated protein sequence. Each sample's HLA type comes from OptiType prediction. We predicted the binding affinity between epitopes and the major histocompatibility complex (MHC) using NetMHC4.[129] Epitopes with binding affinity $\leq$ 500nM, also not present in the wild-type transcript, are reported as neoantigens.

Oncogenicity annotation

OncoKB oncogenicity annotation

OncoKB is an extensive knowledge base that curates the literature on whether specific alterations are either known to be oncogenic or are likely oncogenic.[71] To utilize the most up-to-date annotations, we used the OncoKB API through their Python-based command line interface (https://github.com/oncokb/oncokb-annotator) for point mutations (MafAnnotator.py), copy number alterations (CnaAnnotator.py) and gene fusions (FusionAnnotator.py). To obtain cancer type-specific oncogenicity annotations, we provided the corresponding cancer type for each mutation using standardized OncoTree code names,[130] which only differed from CPTAC code names for three cancer types (LUSC instead of LSCC, HNSC instead of HNSCC, and PAAD instead of PDAC). Somatic point

mutations were annotated based on the genomic change in the MAF file as input. For copy number alterations, OncoKB directly used the GISTIC 2.0 results as input. Lastly, for gene fusions, only the identity of the genes in a fusion pair was provided since OncoKB does not consider breakpoints.

Annotating putative driver missense mutations

As the primary goal of our study was to identify the proteomic correlates of driver alterations and not to identify driver events per se, to explore the consequences of driver events, we defined putative driver alterations by combining two curated knowledge bases (OncoKB[71] and Cancer Gene Census[131]) with computational predictions. We focused on 299 cancer driver genes,[1] supplemented with known oncogenic mutations and computational predictions of driver missense mutations by six methods (Figures S1F–S1H), known mutational hotspots for in-frame indels, and manual curation of less-studied alterations, such as gene fusions. This strategy identified 3,083 somatic point mutations, 2,311 copy number events, and 49 gene fusions as putatively oncogenic for downstream analysis (Figures S1I–S1L; Table S1).

1,289 missense mutations labeled as ''Oncogenic'' or ''Likely Oncogenic'' by OncoKB were considered putative driver mutations. However, because existing annotations like OncoKB often miss rare driver missense mutations, we augmented our analysis with the union of computational predictions from six different methods (CHASMplus tissue-specific and pan-cancer models, BoostDM, CTAT-cancer, CTAT-general, and protein structure hotspot), which have complementary strengths (see section: computational methods to predict driver missense mutations). We added the union of computational predictions because all methods were highly conservative in predicting driver missense mutations (Figure S1F), highly consistent with each other (Figure S1G), and most missense mutations predicted by multiple methods were already annotated by OncoKB as (likely) oncogenic variants (Figure S1H). This resulted in 1,814 candidate driver missense mutations for further analysis. This number (n = 1,814) results from the union of all mutations annotated as ''Oncogenic'' or ''Likely Oncogenic'' in OncoKB (n = 1,289) with any mutation identified by any of the six computational methods (n = 1,699).

Computational methods to predict driver missense mutations

CHASMplus. CHASMplus is a random forest machine learning algorithm that predicts driver missense mutations based on 95 features.[4] Given the larger sample size, a CHASMplus model was trained on data from The Cancer Genome Atlas (TCGA)[132] rather than directly from CPTAC. Because driver missense mutations may be highly cancer type-specific or found at a low frequency across multiple cancer types, predictions from both a cancer type-specific model and a pan-cancer model were used. CHASMplus scores and p-values were obtained through OpenCRAVAT,[133] which can notably score missense mutations even if they were never found in TCGA. Missense mutations were regarded as putative driver missense mutations if the Benjamini-Hochberg FDR was less than 0.01.

BoostDM. We downloaded the BoostDM[5] predictions for almost all possible mutations in 250 gene-tissue-specific pairs from the IntOgen website (www.intogen.org/boostdm). This tool, BoostDM, is a machine learning algorithm that can predict, in a tissue-specific manner, whether a mutation in a cancer gene is likely to be oncogenic or not. We kept only the gene models that matched the cancer types studied in CPTAC and labeled each mutation according to the *boost_dm* flag of the model. This flag is *true* if the model predicts the mutation as oncogenic and *false* otherwise. We then matched the CPTAC mutations to the specific isoform analyzed by BoostDM.

CTAT-cancer, CTAT-general, and protein structure hotspots. The results of three driver missense methods (CTAT-cancer, CTAT-general, and protein structure hotspots) were obtained from our previous analysis of TCGA.[1] CTAT-population is a consensus of four machine learning methods (SIFT,[134] PolyPhen2,[118] VEST,[135] and MutationAssessor[136]) that predict the pathogenicity of missense mutations by scoring each variant using a PCA-based strategy. A drawback of CTAT-general was that it was not tailored for analyzing somatic mutations. To overcome this weakness, using a similar PCA-based strategy, CTAT-cancer is a consensus of four machine learning methods (CanDrA,[137] fathmm,[138] CHASM,[139] and TransFIC[140]) that predicts the oncogenicity of missense mutations, rather than general pathogenicity. Lastly, structural clustering is a consensus of at least two out of four clustering algorithms (HotMAPS,[10] HotSpot3D,[141] 3DHotSpot.org,[142] and e-Driver3D[9]) that identify mutations that cluster in 3D protein structures.

Annotating putative driver in-frame indels

To overcome the limited number of computational approaches to predict whether an in-frame indel is likely a cancer driver mutation, we took a multi-stage strategy. Similar to missense mutations, we first included all ''Oncogenic'' or ''Likely Oncogenic'' mutations from OncoKB as putative driver mutations. We also included any in-frame indels in mutational hotspots from Chang et al.[143] The first two steps resulted in the annotation of 89 in-frame indels as putative drivers. Lastly, we manually reviewed the literature for 176 in-frame indels that were found in 299 previously implicated driver genes, which resulted in the rescue of 16 additional putative drivers in frame indels based on proximity to similar mutations known to have a functional effect in experimental assays (Table S1).

Annotating driver putative Loss-of-Function (pLOF) mutations

Putative Loss-of-Function mutations (including nonsense, frameshift indels, translation start site, and stop-loss mutations) are commonly regarded as driver mutations if they occur in known tumor suppressor genes.[144] OncoKB utilizes this rule of thumb in a cancer-type-agnostic manner for annotating pLOF mutations as oncogenic.[71] However, we found such a strategy likely resulted in elevated false positive driver mutation calls, as evidenced by a substantial correlation between gene length and the frequency of pLOF mutations (Figure S1D). We, therefore, annotated pLOF mutations based on cancer type-specific annotations of TSGs using the 20/20+ method.[145] Briefly, we regarded a driver gene as a TSG for a cancer type if it was at least nominally statistically significant for having a high TSG score (p-value<0.05) based on data from TCGA.[1] Notably, the cancer type-specific TSG annotations eliminated the correlation between gene length and pLOF mutation frequency (Figure S1I), suggesting that most false positives were removed.

Annotating driver copy number alterations

Copy Number Alterations (CNA) were summarized at the gene level through GISTIC 2.0, with only ''high-level'' amplifications and deletions used for further analysis, which already adjusts for broad arm-level differences that may not be specific to the gene. Driver CNAs were defined fromtwo sources (OncoKB and the Cancer Gene Census) that consider whether the directionality of the CNA (amplification vs. deletion) is consistent with the gene's known role in tumorigenesis (oncogene vs. tumor suppressor gene, respectively). For OncoKB, only CNAs annotated as ''Oncogenic'' were considered drivers. For Cancer Gene Census,[131] we regarded any high-level amplification/deletion ina CGC gene annotated as appropriate for amplification/deletion, respectively, as a putative driver CNA. Because both annotations from OncoKB and CGC were highly consistent (Figure S1K), we regarded a CNA in a gene as a putative driver if either of the approaches indicated it.

Driver annotation of gene fusions

We regarded fusions annotated as ''Oncogenic'' or ''Likely Oncogenic'' fusions from OncoKB as putative drivers. To reduce the potential for passenger fusion events being erroneously labeled as a driver, we only used OncoKB annotations where the precise gene fusion that contained both genes was known to be likely/oncogenic rather than imprecisely defined gene fusions (e.g., ALK fusions without a defined fusion partner). We then manually reviewed cases to rescue oncogenic fusions where a known oncogene driven by fusion events may have had a non-canonical fusion partner. Given the importance of kinase fusions in cancer,[124] we focused on kinase fusions that resulted in an in-frame fusion event with the kinase domain intact. To do this, we used agfusion to annotate the consequence of gene fusions on the protein,[146] including in-frame status and protein domains that were lost or retained. Notably, as expected, fusions containing a known oncogene were enriched for in-frame fusion events (Figure S1L), which was not the case for tumor suppressor genes. This resulted in the inclusion of 49 fusion events as putative drivers across CPTAC.

Normalization of protein abundance by RNA expression

To analyze the *cis*-effects that happen specifically at the protein level rather than iption level, we first performed a linear regression between RNA expression ($\log_2$(FPKM+1)) and protein abundance of each gene using the Python statsmodels package. We then computed the Pearson residual for each sample, which normalizes the difference between the observed protein abundance and expected value based on RNA expression (residual) by their standard deviation. The *cis*-effect analysis (see section: Genetic driver alterations' *cis* impact on RNA, proteome, and PTMs) was then repeated using the Pearson residual rather than protein abundance directly.

Calculation of relative solvent accessibility of amino acid residues

Relative Solvent Accessibility (RSA) is a continuous score to characterize whether an amino acid is buried vs. accessible on the protein surface. RSA is calculated as the fraction of the surface area of an amino acid residue that is solvent-accessible based on an available protein structure. To systematically calculate RSA across the proteome, we used predicted protein structures from AlphaFold (version 1, downloaded from https://alphafold.ebi.ac.uk/).[147] Using default parameters, RSA values were computed using DSSP[148] via the BioPython API using default parameters. Notably, as RSA depends on normalizing the accessible surface area by a theoretical maximum for each residue, the BioPython implementation uses the approach by Sander and Rost.[149]

Protein-protein interaction databases

We integrated protein-protein interactions from the following sources: STRING[150] (minimum score above 700), MINT,[151] BioGrid,[152] IntAct,[153] CORUM[154] and the unbiased experimental interactome.[155] While the union of all these databases had 710,748 interactions, for the rest of the analyses, we only kept those interactions described in at least three databases (n=14,768).

Driver mutation interaction term regression model for protein-protein interactions

As outlined previously, protein-protein interactions supported by multiple databases (see above) displayed higher protein co-expression than compared to random protein pairs. To statistically test whether putative driver mutations may alter protein co-expression of the driver gene (g) with a known PPI partner, we used a linear regression model with a non-linear interaction term. Interaction terms are a widely used technique to capture when the relationship between the outcome (e.g., protein abundance of a PPI partner) and an independent variable (e.g., protein abundance of a cancer driver gene) depends on a third variable (e.g., mutation status).[156] After including additional covariates, this resulted in the following regression model to learn the β coefficients:

$$Y = \beta_0 + \beta_1 X_g + \beta_2 M_g + \beta_3 (X_g * M_g) + \beta_4 P + \beta_5 C + \epsilon \quad \text{(Equation 1)}$$

where Y is a (n x 1) vector representing the protein abundance of the PPI partner, X is a vector representing the protein abundance of the cancer driver gene of interest (g), M is a binary vector indicating the driver mutation status for that particular driver gene (g) in a tumor sample, (X*M) is the interaction term composed of the element-wise product of X and M, and the tumor purity (P) calculated by ESTIMATE[127] was also included as a covariate. For pan-cancer analyses, cancer type was one-hot encoded and was included as an additional covariate (C). Lastly, the error ε is assumed to be normally distributed with a constant variance σ.

To assess for non-zero β coefficients, indicating a potential *trans*-effect of a driver mutation on a PPI partner, we performed a Wald test for β_2 and β_3, which resulted in p-values P_2 and P_3. To summarize the overall impact of the driver mutation, we then used Fisher's

method to obtain a single combined p-value. Fisher's method assumes the log-transformed p-values follow a chi-squared distribution with 2*(number of hypotheses, |K|) degrees of freedom.

$$\chi^2_{2|K|} \sim -2\sum_{k\in K} log(P_k) = -2\,log(P_2) - 2\,log(P_3) \quad \text{(Equation 2)}$$

This procedure was applied for all genes with at least five putative driver mutations and all PPI partners that are supported by at least three databases. Only proteins with less than 50% missing values were tested. The Benjamin-Hochberg false discovery rate (FDR) approach was then applied, and all events with a q-value less than 0.1 were deemed statistically significant. Notably, repeating the same analysis using RNA instead of protein abundance levels, we could only recover approximately half of all correlations between interacting pairs (~52% on average across the ten cancer types), ranging from 44% in BRCA to 73% in COAD.

Robustness and subsampling analysis of the interaction term regression model for protein-protein interactions

While including a product interaction term in a regression model is well established, at least two factors could adversely impact the proper control of the false discovery rate for our data in practice. First, p-values from the Wald test may not be independent and therefore lead to an inflated p-value estimate when combined using Fisher's method. Secondly, protein abundance outliers may have high leverage in the regression model and potentially drive spurious associations. To assess the robustness of the interaction term regression model, we randomly permuted the samples labeled as driver mutant (M vector) in our pan-cancer analysis driver mutant labeled samples permuted, and all statistically significant *trans*-effects would be false positives. Consistent with the interaction term regression model controlling the Type I error, QQ plots show that the permuted labels closely match the null hypothesis of a uniform p-value distribution (Figure S3I). In contrast, on the observed (unpermuted) data, the interaction term regression model identifies events with p-values much lower than expected based on the null hypothesis. We then repeated the permutation 1,000 times to generate a distribution of the number of significant events (Figure S3J). We found that the mean number of false positives from the permutation analysis closely matched the estimated FDR threshold (~4 mean false positives, 51 events as statistically significant). This suggests that the interaction term regression model was appropriately controlling the Type I error rate.

We next performed a subsampling analysis to identify how sample size impacted our PanCan analysis. To do this, we randomly subsampled tumors without replacement for sizes ranging from 2% to 100% of the full data. For each subsampled dataset, we repeated the analysis described in the section interaction term regression model for protein-protein interactions. After 100 subsampled analyses, we found a roughly linear increase in the number of significant events as the sample size increased (Figure S3K).

Copy number mediation analysis

Because copy number events may impact more than one gene, a central question is whether the protein product of a particular gene might mediate the downstream impact on a phenotype. To address this question, we posed the problem in terms of a mediation analysis where the copy number status of an arbitrary Gene A within the CNA is assessed for impact on the expression level of Protein A, which would then have a *trans*-impact on the expression level of Protein B (Figure 3F). To restrain testing all potential pairs of A and B, we restrict our analysis to those with protein interaction support from at least three databases (see the section titled "Protein-protein interaction databases"). In causal inference theory, mediation analysis is typically formulated in terms of exposure (X), mediator (M), covariates (C), and outcome (Y). For our application, X is a binary variable indicating a putative oncogenic CNA event for Gene A, M is the expression level of Protein A, C is a one-hot encoding of the cancer type, and Y is a vector of the expression level of Protein B. One commonly used statistical approach to reject the null hypothesis of no mediation is the product-of-coefficients method,[157] which leverages the following two regression equations:

$$\widehat{M} = \widehat{h_1} + \widehat{\alpha}X + \widehat{\zeta_1}\,C \quad \text{(Equation 3)}$$

$$\widehat{Y} = \widehat{h_2} + \widehat{\tau}X + \widehat{\beta}\,M + \widehat{\zeta_2}\,C \quad \text{(Equation 4)}$$

where $\widehat{h_{1-2}}$ are intercept terms, $\widehat{\zeta_{1-2}}$ are coefficients for the covariates, and $\widehat{\alpha}$ and $\widehat{\beta}$ are the estimated coefficients of interest. In the product-of-coefficients method, the product of the two terms ($\widehat{\alpha}\widehat{\beta}$) is then tested for being non-zero. This is usually done by repeated bootstrapping or dividing by the standard error of $\widehat{\alpha}\widehat{\beta}$ by a first-order approximation and comparison to a standard normal distribution.[157] A more recent approach to mediation analysis is to use a natural effect model that reformulates Equation two (e.g., Equation 4) in terms of counterfactuals. For the CNA mediation analysis, we, therefore, used a natural effect model in the MedFlex R package using the imputation-based approach,[158] which, unlike the product-of-coefficient method, handles missing values in the outcome variable. To assess whether the null hypothesis of no mediation could be rejected, we used a Wald test based on the standard error estimate by the sandwich estimator.

Phosphorylation interaction term regression model for protein-protein interactions

We also statistically tested whether phosphorylation levels at protein interfaces may influence protein co-expression of the phosphorylated protein (p) with a known PPI partner. We first identified phosphorylation sites at interaction domains on proteins using

3DMapper, a tool that locates protein positions on protein interfaces (https://github.com/vicruiser/3Dmapper). Cross-referencing these interface-mapped sites with known protein regulatory phosphorylation sites documented in PhosphoSitePlus[159] revealed that 58 of the 212 total interface phosphorylation sites detected by 3DMapper are, indeed, involved in regulatory mechanisms according to PhosphoSitePlus annotations. To binarize the extent of phosphorylation levels for more straightforward testing, we converted interface phosphorylation levels to 1 for patients with phosphorylation levels in the upper quartile, 0 for patients with phosphorylation levels in the middle 50% of measures at the site, and -1 for patients with lower quartile phosphorylation levels. We used a linear regression model with a non-linear interaction term, similar to what was implemented for the driver mutation PPI analysis, to ask whether patients with high phosphorylation at a given site exhibited different protein-protein correlations from patients with intermediate or low phosphorylation levels at the same site. After including additional covariates, this resulted in the following regression model to learn the β coefficients:

$$Y = \beta_0 + \beta_1 X_p + \beta_2 M_p + \beta_3 (X_g * Ph_p) + \beta_4 P + \beta_5 C + \epsilon \quad \text{(Equation 5)}$$

where Y is a (n x 1) vector representing the protein abundance of the PPI partner, X is a vector representing the protein abundance of the phosphorylated cancer driver of interest (p), Ph is a binary vector indicating the binary-converted phosphorylation levels for a given interface site in that particular driver protein (p) in a tumor sample, (X*Ph) is the interaction term composed of the element-wise product of X and Ph, and the tumor purity (P) calculated by ESTIMATE[127] was also included as a covariate. For pan-cancer analyses, cancer type was one-hot encoded and was included as an additional covariate (C). Lastly, the error ε is assumed to be normally distributed with a constant variance σ.

Regression analysis of neoantigens and inferred immune infiltration

To analyze factors that may influence the immunogenicity of neoantigens, we examined the correlation of the number of neoantigens ($\log_2$ transformed) within a tumor (neoantigen burden) with inferred T-cell infiltration. Two methods for inferring CD8 T-cell infiltration were tested, CIBERSORTx absolute[160] and a cytotoxic t lymphocyte (CTL) score composed of the average expression of several marker genes (CD8A, CD8B, GZMA, GZMB, and PRF1).[161] At the pan-cancer level controlling for cancer type, both methods were significantly correlated with neoantigen burden (CIBERSORT absolute CD8 T-cell: p-value=0.002, CTL: p-value=4e-4; Wald test), with four individual cancer types (COAD, UCEC, LUAD, and BRCA) being nominally significant in both analyses. To analyze whether driver alterations may alter the correlation between neoantigen burden and inferred T-cell infiltration, we used a linear regression model with an interaction term between driver alteration status and neoantigen burden and inferred T-cell infiltration to analyze whether driver alterations may alter the correlation between neoantigen burden and inferred T-cell infiltration. Like the interaction term regression model in the section "Interaction term regression model for protein-protein interactions", we assessed statistical significance using a Wald test (q-value<0.1). To assess whether consideration of the protein / RNA expression of a gene improves the correlation of neoantigen burden with CTL, we recalculated neoantigen burden based on predicted neoantigens found in genes with at least certain levels of RNA or protein abundance in each tumor. To assess whether this improved the correlation with CTL, we performed a likelihood ratio test that compared a model with the original neoantigen burden feature to a model that also included expression-filtered neoantigen burden.

Analysis of *RB1* oncogenic alterations

To further analyze the impact of oncogenic alterations in *RB1*, we performed three downstream analyses on expression data (RNA-seq and proteomics), ChIP-seq data[162–164] and crispr screens (DepMAP). First, we performed differential expression analysis of putatively oncogenic alterations in *RB1*, either point mutations or CNAs, versus wild-type by using a linear regression model that controls for tumor type and tumor purity (ESTIMATE). A Wald test assessed statistical significance at a q-value<0.1. Differential expression was done separately for RNA-seq (based on $\log_2$ FPKM values) and proteomics. Secondly, given that RB1 is known to aid transcriptional repression of target genes,[56] we next examined RB1 ChIP-seq from three different studies.[162–164] We downloaded uniformly processed ChIP-seq coverage files from CistromeDB (ids: 101666, 1872, and 49416),[165] as well as regulatory potential (RP) scores[166] that describe how close ChIP-seq peaks are to the transcriptional start site of a gene. We then normalized the RP scores based on the maximum RP observed in each dataset. Lastly, we analyzed which genes might be preferentially essential in *RB1*-altered cell lines from large-scale CRISPR screens. We downloaded quantified essentiality scores and mutation calls from DepMAP.[60] Given the absence of CNA calls for deletions, we only used point mutations that were putatively loss-of-function (i.e., nonsense, frameshift indels, etc.) in RB1. Using a linear regression model that adjusts for cell lineage, we then used a Wald test to assess whether RB1 alteration status was associated with differential gene dependencies (q-value<0.1).

Drug connectivity

Differentially Expressed Proteins (DEPs) were identified using Limma (trend=True) between various clinical, driver gene mutation status, and pan-cancer proteogenomic groups after filtering for proteins with adjusted p-value<0.05. The identified DEPs were then filtered for gene symbols measured in the L1000 assay (978 landmark genes + 9122 Best Inferred Genes). These driver-gene-specific signatures were input to calculate normalized weighted connectivity scores (WTCS) against the Library of Integrated Network-Based Cellular Signatures (LINCS) L1000 perturbation-response signatures. The scores were computed using the sig_queryl1k_tool pipeline (https://hub.docker.com/u/cmap) and the LINCS L1000 Level 5 compound (trt_cp) signatures from CLUE (https://clue.io, "2021

Release''). The resulting normalized connectivity scores were summarized across cell lines using the maximum quantile of all the scores of the same compound (Qhi=67, Qlow=33) as previously described.[57] The corresponding compound metadata were obtained from CLUE (clue.io, ''Expanded CMap LINCS Resource 2020 Release'') and were used to filter and identify compounds with existing clinical annotations (have been evaluated in at least a Phase I trial). The normalized drug connectivity scores of each driver gene signature were then used to calculate individual MOA and Drug Target enrichment scores by applying the Gene Set Enrichment Analysis of the ''fgsea'' R package and by utilizing the compounds' respective MOA and Drug Target Annotations from CLUE.

Gene Set Enrichment Analysis (GSEA) on signature gene coefficients

We submitted the signature gene coefficients to Gene Set Enrichment Analysis (GSEA) on a subset of the GO terms that reduced redundancy between pathways while retaining the gene coverage.[167] The MSigDB c5 v7.4 gene set was first filtered to remove terms with less than 15 or more than 200 genes, and reduction was carried out using the authors' implementation (https://github.com/RuthStoney/set-cover-and-set-packing-to-reduce-redundancy-in-pathway-data) with unweighted coverage, resulting in a reduced set of 698 terms. After GSEA on the reduced set, 633 terms with enrichment scores were clustered based on term semantics using the SimplifyEnrichment R package (Gu and Hübschmann, 2021), and the term groups were manually annotated.

The Kinase Library enrichment analysis

Full description of the substrate specificities atlas of the Ser/Thr kinome can be found in the KL paper.[109] The phosphorylation sites detected in this study were scored by all the characterized kinases (303 S/T kinases), and their ranks in the known phosphoproteome score distribution were determined as described above (percentile score). For every non-duplicate, singly phosphorylated site, kinases ranked within the top 15 kinases for the S/T kinases were considered biochemically predicted kinases for that phosphorylation site. Towards assessing a kinase motif enrichment, we compared the percentage of phosphorylation sites for which each kinase was predicted among the downregulated/upregulated phosphorylation sites (sites with a q-value of 0.1 or below) versus the percentage of biochemically favored phosphorylation sites for that kinase within the set of unregulated sites in this study (sites with q-value above 0.1). Contingency tables were corrected using Haldane correction (adding 0.5 to the cases with zero in one of the counts). Statistical significance was determined using a one-sided Fisher's exact test, and the corresponding p-values were adjusted using the Benjamini-Hochberg procedure. Then, for every kinase, the most significant enrichment side (upregulated or downregulated) was selected based on the adjusted p-value and presented in the volcano plots and bubble maps. In the volcano plots, significant kinases (q-value<=0.1) for both upregulated and downregulated analysis were plotted on both sides. Bubble maps were generated with size and color strength representing the q-values and frequency factors, respectively, only displaying significant kinases (q-value<=0.1). Significant kinases (q-value<=0.1) for both upregulated and downregulated analysis were plotted using the parameters of the more significant side. For example, as shown in Figure 5I, immune signaling in the protein-based pathway analysis matched the activity of known immune-related kinases, such as MAPKAPK2/3/5, IKKB/E, and TBK1. Hypoxia was enriched at the protein level only in ccRCC, similar to the activity pattern of the stress-related kinases PAK1-6.

Genetic driver alterations' *cis* impact on RNA, proteome, and PTMs

Cis-effects are here defined to be the effect of mutations in a particular cancer driver gene on the levels of RNA, protein, and phosphorylation corresponding to the same gene. To assess the impact of cancer driver mutations on RNA, protein, and phosphorylation, scores were calculated for each driver event for *cis* effects. The *cis*-effect scores were calculated using the outputs of a linear regression modeling the effects of a particular driver gene mutation status on the protein abundance, RNA expression, or phosphorylation corresponding to the gene.

$$Y = \beta_0 + \beta_1 M_g + \beta_2 P + \beta_3 C + \epsilon \quad \text{(Equation 6)}$$

where Y denotes an (n x 1) vector containing either the protein abundance, the RNA expression, or phosphoproteomic level at a residue pertaining to the driver gene, M is a binary vector indicating the driver mutation status for that particular driver gene (g) in a tumor sample, and the tumor purity (P) calculated by ESTIMATE[127] was also included as a covariate. For pan-cancer analyses, cancer type was one-hot encoded and was included as additional covariates C. The error ε is assumed to be normally distributed with a constant variance σ.

After running the model for each driver and determining the associated correlation and Benjamini-Hochberg FDR adjusted p-values within each cancer type at each level of omics data, the *cis*-effect score was determined by multiplying the -$\log_{10}$(p-adj) by the sign of the correlation coefficient between the mutated driver and the effect on the omics. This retains the directionality of the mutated driver on the target of interest, relative to samples that are wild-type for that cancer driver gene. This procedure was conducted broadly on the pan-cancer level as well as in a tissue-specific manner. The top results for the pan-cancer *cis*-effects and the cohort-specific results are depicted in Figure 2B. In the case of phosphoproteomics, we kept the absolute highest score for each gene.

Similarity of drivers through *trans*-effects on proteome and PTMs

As opposed to *cis*-effects, *trans*-effects denote the impact that driver mutations have on RNA, protein, and phosphorylation levels not corresponding to that of the mutated driver gene. *Trans*-effects were determined in this study by running a linear regression for each

cancer driver, similar to the process described above for the *cis*-effect score determination and the *trans*-effect regression model for protein-protein interactions. For each cancer driver, all *trans* protein levels and PTM were tested with the following model:

$$Y = \beta_0 + \beta_1 M_g + \beta_2 P + \beta_3 C + \epsilon \quad \text{(Equation 7)}$$

where Y is a (n x 1) vector representing the protein abundance of a given protein or the phosphorylation levels of a specific residue, M is a binary vector indicating the driver mutation status for that particular driver gene (g) in a tumor sample, and the tumor purity (P) calculated by ESTIMATE[127] was also included as a covariate. For pan-cancer analyses, cancer type was one-hot encoded and was included as an additional covariate (C). Lastly, the error ε is assumed to be normally distributed with a constant variance σ.

For each driver, scores were assigned to all corresponding *trans* proteomic and phosphorylation events by multiplying the -$\log_{10}$(p) of the association with driver mutation by the sign of the coefficient, as was conducted for the *cis*-effect scores. To determine cancer drivers that are associated with similar proteomic and phosphorylation changes when mutated, a ''driverness similarity'' calculation was conducted for each pair of drivers. The scores of all the drivers' events were compared to determine a holistic correlation between the two drivers, and the *trans*-events with scores more extreme than 5 or -5 for both drivers were highlighted (Figure 4A). The distribution of similarity scores for all driver pairs is shown in Figure 4B.

Multi-omic signatures and clusterings

In this pan-cancer cohort, shared data types available for clustering were whole transcriptome RNA-seq, proteome, and phosphoproteome. To harmonize the RNA data with the normally distributed proteomic data, we applied a normal inverse transformation, median-centered the data, and stabilized using median absolute deviation. We subsetted for samples that had matched RNA, protein, and phosphoprotein in the entire dataset and concatenated these three matrices. To reduce the transcriptome space to a comparable feature size to the intersected proteome data, we selected the top 5,000 highly variable genes, ranked by the coefficient of variation. The combined matrix contained gene expression for 5,000 highly variable mRNA genes, 5,716 proteins, and 3,341 phosphoproteins. The description of *SignatureAnalyzer* conveyed the following details: (i) The overall rationale for the tool; (ii) Examples of papers where the tool was successfully applied; (iii) The utilization of the tool in this paper; and (iv) Analysis of the robustness of signatures concerning the cohort size which was provided in Geffen et al.[168]

C3PO

To explore the overall contribution of genomic variants to protein variability within cancer types, we built a combined polygenetic protein abundance prediction algorithm for oncology, called C3PO. C3PO is largely based on polygenic risk scores (PRS),[62–64] a mathematical approach typically used to provide an additive germline polygenic score that correlates with disease risk in an individual. In other words, the PRS for any individual sample is calculated by summing the genome-wide genotypes, weighted by that genotype's effect size estimate. This can be represented by the following equation,

$$PRS_i = \Sigma G_{ij} S_j \quad \text{(Equation 8)}$$

where *PRSi* is the polygenic risk score for a single sample *i*; *G* is the genotype for the *i*-th sample at genomic position *j* (and can take the values of 0, 1, or 2), multiplied by S_j the effect size that G_j contributes to disease. Thus, generating a summative sample-level risk profile score of disease for that sample.

Building on the PRS framework, we sought to determine the polygenic abundance prediction score for all of the proteins measured in the proteome. A simple mathematic model to represent this step would be as follows:

$$C3PO_{ik} = \Sigma G_{ij} S_{jk} \quad \text{(Equation 9)}$$

where a *C3PO score* is calculated for each sample *i* and protein *k* as the sum of any somatic mutations *G* in a gene-network *j* weighted by G_j's effect size (*S*) on protein *k*. Here G_j can take values of 0 or 1 as somatically mutated or not mutated. The range of *S* takes the normalized units of protein abundance calculated by the CPTAC consortia. See above. Before moving on it is sufficient to expound on the methodology to calculate both G and S.

The concept of somatic mutations *G* across many samples breaks as a viable additive model because of the issue of a right-skewed tail of somatic mutations in cancer genomes (Figure S1J). In other words, unlike common germline mutations where many samples share the same germline mutations, shared somatic mutations are infrequent. To overcome this issue of rare mutations, as is often performed with rare mutation analysis, we aggregated mutations into burden units based on NeSTed protein systems according to Zheng et al.[65] (Step 1, Figure S7A, variable *G* in equation). Briefly, this hierarchical nested network was generated by integrating spatial proximity measures from aggregate spatial transcriptomic and co-immunoprecipitation assays to generate nests of genes that work in concert within the cell. Doing so provided broader coverage of the possible mutational impacts on protein (Figure S7A).

The second piece, effect size (*S*), of the standard PRS calculation was also altered in C3PO calculations. These are typically generated from population-based case-control datasets using genome-wide association studies (GWAS). Here, instead of a GWAS model of cases and controls, C3PO estimates protein abundance effect sizes. More specifically, we used the Cohen's D difference between protein levels in altered vs. wild-type tumor samples from each cancer type and across all samples in CPTAC, i.e., pan-cancer measures (Step 2, Figure S7B, variable *S* in equation). Cohen's D is generated by comparing samples with mutations in a NeSTed protein system to those without mutations.

Step 3 in our theoretical C3PO model was, to sum up our estimates to build a protein prediction score, or C3PO-score, for every protein (*k*) in each sample (*i*). To perform this task, we built two matrices based on the previous two steps. We built these matrices to avoid using computationally expensive for-loops. Our first matrix captured genetically mutated NeSTs for each sample, i.e., a 0,1 matrix (Sample x Mutated-NeSTs). We then generated a second matrix capturing the effect sizes (Cohen's D statistics) for each NeST's effect on protein abundance (NeST x Protein-abundance). In other words, we generated a weighted matrix that could be used to supply the weights to our C3PO score. Only weights with p-values less than 0.1 were included (Step 3, Figure S7A). This allowed for the aggregation of genetic alterations with smaller effects on protein levels to influence our model. Finally, the two observed matrices were multiplied to produce a global proteomic prediction for each sample, i.e., a sample by protein matrix of C3PO scores. This exercise was repeated for each substrate (DNA, CNA amplification, and CNA deletion) and ultimately combined into a cumulative CP3O prediction. Despite gross overtraining of this model, we estimated the Pearson correlation of C3PO and measured correlation coefficients for each substrate individually (Figure S7B) and collectively (Figure S7C). The explained variance reported in the manuscript was calculated by taking the mean floor and mean ceiling of each C3PO substrate prediction. The trained weight matrix from CPTAC was applied to somatic alterations from CCLE and CPTAC retrospective samples not included in this analysis.

Enrichment analysis was performed to associate high smoking scores with high C3PO scores in the chromatin modification pathway for LUAD samples. This was performed by taking the root-mean-squared of the smoking scores and chromatin modification scores generated by C3PO. Samples were classified as high if scaled scores were greater than zero and low if scores were less than or equal to zero. Samples were placed in a 2x2 table for comparison using Pearson's chi-squared test with Yates' continuity correction. C3PO is available at https://github.com/MHBailey/C3PO_polygenic.

Proteomics data processing

The details are provided in Geffen et al.[168]

QUANTIFICATION AND STATISTICAL ANALYSIS

RNA and protein quantification

The process of RNA data processing and quantification, as well as proteome quantification, has been outlined in the sections titled "RNA data processing and quantification" and "Proteomics data processing", respectively. The statistical analysis methodology and its corresponding details can be found both within the main text and in the relevant sections of the STAR Methods.

Gene expression, proteomic, and phosphoproteomic feature-associated markers

Proteogenomic data was used to perform pairwise differential analysis between groups of samples. We performed the T-test, Wilcoxon rank-sum test, and limma-trend test to do the differential expression analysis for protein, phosphoprotein, and gene expression between the groups with and without a given feature at the cohort-specific level and pan-cancer level (adjusted by cancer type). This analysis requires at least five samples in each group with non-missing values. P-values were adjusted using the Benjamini-Hochberg FDR method.

Statistical power analysis of driver gene discovery

We compared the statistical power to identify genes with a significantly elevated mutation rate ("significantly mutated genes") for cancer types in common between The Cancer Genome Atlas (TCGA) and CPTAC. The statistical power calculations used the cancerSeqStudy R package (https://github.com/KarchinLab/cancerSeqStudy).[145] Similar to previous power calculations,[1,169] we assumed that the number of gene mutations followed a binomial distribution. The background mutation rate parameter (p=mutations per base) was estimated for each cancer type by calculating the median exome-wide mutation rate across tumor samples, assuming for simplicity ~30,000,000 nucleotide bases in coding regions as potentially mutable. In line with our previous statistical power analyses,[1,169] a weighting factor (default parameter) was used to correct for variability in mutation rate across genes and that not all mutations can generate a non-silent mutation. A gene was considered statistically significant at a family-wise error rate of 0.1. Calculations of the sample size required to achieve 90% statistical power were carried out for various mutation frequencies above the background mutation rate, including 20%, 10%, 5%, and 2% of mutated tumor samples above the background mutation rate (i.e., effect size=0.2, 0.1, 0.05, and 0.02, respectively) (Figure S1M).

ADDITIONAL RESOURCES

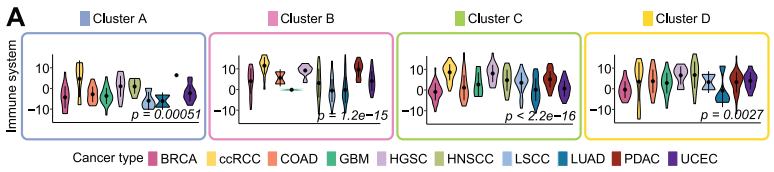
Comprehensive information about the CPTAC program, including program initiatives, investigators, and datasets, are available at the CPTAC program website: https://proteomics.cancer.gov/programs/cptac.

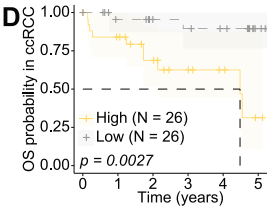
For the Pan-Cancer proteogenomics collection papers, along with links to the data and supplementary materials associated with these publications, please visit the Proteomic Data Commons (PDC) at https://pdc.cancer.gov/pdc/cptac-pancancer.

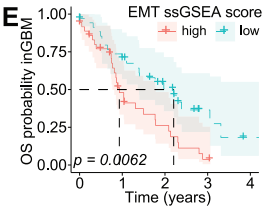
Cell
Article

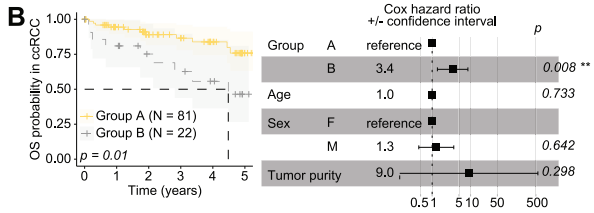
CellPress
OPEN ACCESS

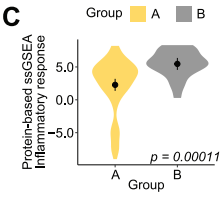
Supplemental figures

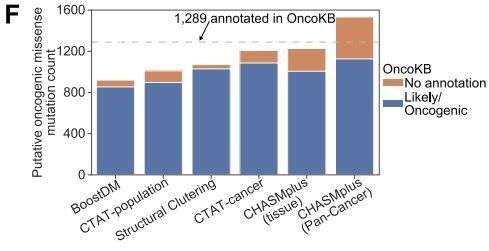


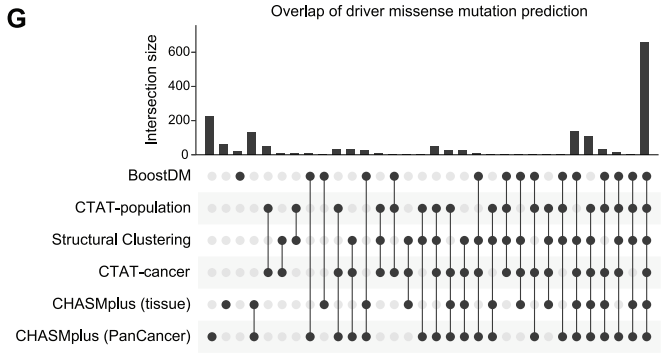


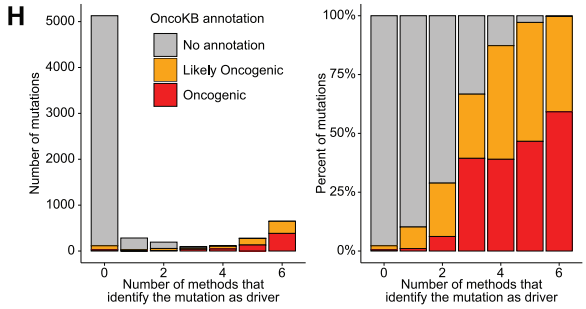


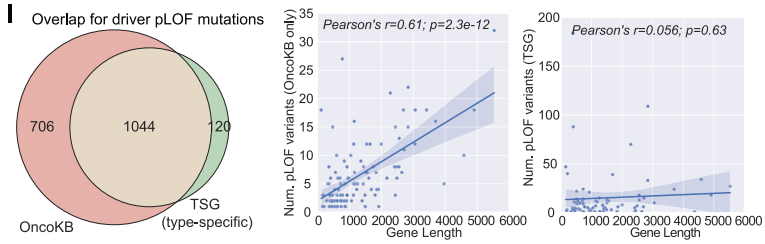


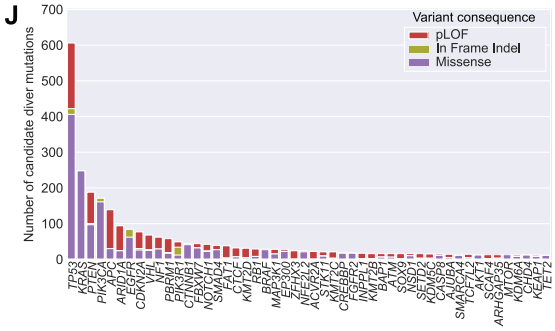


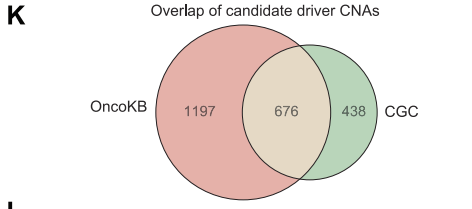


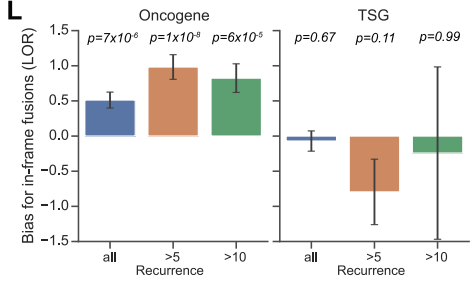


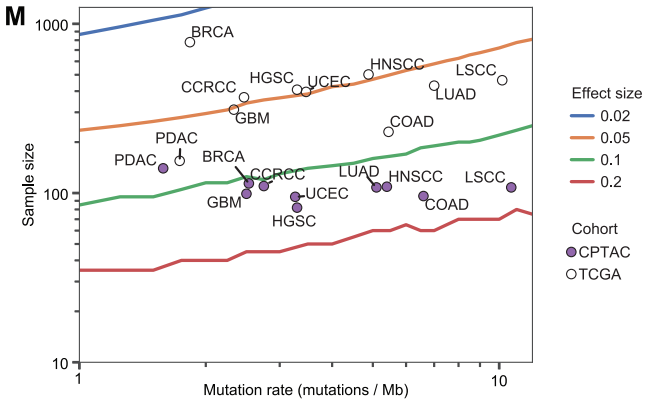












(legend on next page)

Figure S1. Driver mutation annotation process and multi-omic clusters, related to Figure 1

(A) Violin plots showing the cancer-type pathway enrichment score distributions derived from single-sample gene set enrichment analysis (ssGSEA) on the proteomic data in the immune system pathway, grouped by multi-omic clusters. One-way ANOVA p values are shown, and data are represented as mean ± CI.

(B) Overall survival (i.e., OS) difference between group A (i.e., clusters C and D) and group B (i.e., clusters A and B as the counterpart) in ccRCC indicated by Kaplan-Meier plot. Log-rank test p = 0.01. A Cox proportional hazards model p value of 0.008 is attained after adjusting for age, sex, and tumor purity. Whiskers represent ± 95% confidence intervals of the hazard ratios.

(C) The violin plots compare the distribution of protein-based ssGSEA score (i.e., inflammatory response hallmark pathway) between the two ccRCC groups depicted in B). A Wilcoxon signed-rank test p value of 0.00011 is attained. Data are represented as mean ± CI.

(D) Overall survival (i.e., OS) difference between the high group (inflammatory response ssGSEA upper quartile) and the low group (e.g., lower quartile) of the entire ccRCC cohort indicated by the Kaplan-Meier plot. Log-rank test p value = 0.0027. A Cox proportional hazards model p value of 0.011 is attained after adjusting for age, sex, and tumor purity. High, inflammatory response ssGSEA upper quartile; low, inflammatory response ssGSEA lower quartile.

(E) Kaplan-Meier plot indicating the association between overall survival and EMT ssGSEA score expression in the CPTAC GBM cohort. The high EMT ssGSEA score (upper quartile of the cohort) is significantly associated with poor survival, as indicated by the Log-rank test p value.

(F) Bar plot showing the number of predicted driver missense mutations by each of six methods, which is further stratified by whether the mutation was previously known (blue) or not known (orange) to be oncogenic by OncoKB.

(G) Upset plot indicating the overlap in driver missense mutation predictions between computational methods.

(H) Bar plot showing the relationship between the number of computational methods predicting a missense mutation to be oncogenic and its support in OncoKB. We used all six computational methods (STAR Methods) independently to classify the missense mutations and then tallied each mutation according to how many independent callers identified it. The x axis indicates how many of these six methods identified an individual mutation as a putative driver. Missense mutations called by more computational methods include a higher proportion of driver aberrations. The left panel shows the absolute numbers of mutations, and the right panel presents the percentages of mutations in each category.

(I) Left, Venn diagram indicating overlap between the labeling of putative loss-of-function (pLOF) variants as oncogenic by either OncoKB (cancer type-agnostic) or by tumor suppressor gene (TSG) classifications by the TCGA PancanAtlas (cancer-type-specific). Right, the correlation between gene length and frequency of pLOF variants for the two pLOF oncogenicity annotation approaches. The shaded area represents 95% confidence intervals.

(J) Bar plot showing the frequency of annotated driver mutations by gene, stratified by the type of mutation (pLOF, in-frame indel, or missense mutation).

(K) Venn diagram showing the overlap between oncogenicity annotations for copy-number alterations (CNAs) between OncoKB and the Cancer Gene Census (CGC).

(L) Bar plot showing the log-odds ratio for gene fusions resulting in in-frame events, stratified by the type of driver gene (oncogene or tumor suppressor, TSG) and by frequency of fusion event. Error bars represent ±1 SEM.

(M) Statistical power for detecting cancer driver genes at defined fractions of tumor samples above the background mutation rate (effect size with 90% power) is depicted. Circles indicate the ten cancer types in CPTAC and TCGA placed according to the study sample size and median background mutation rate.

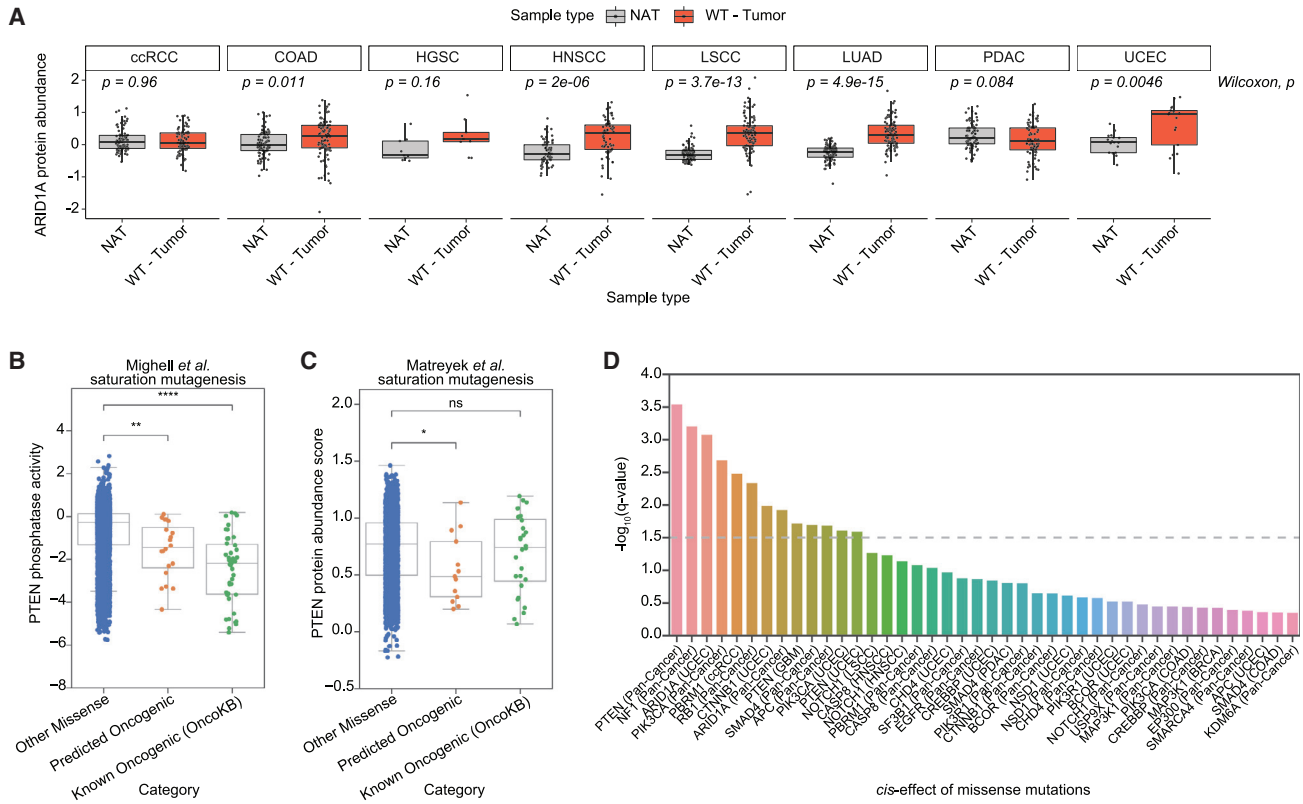


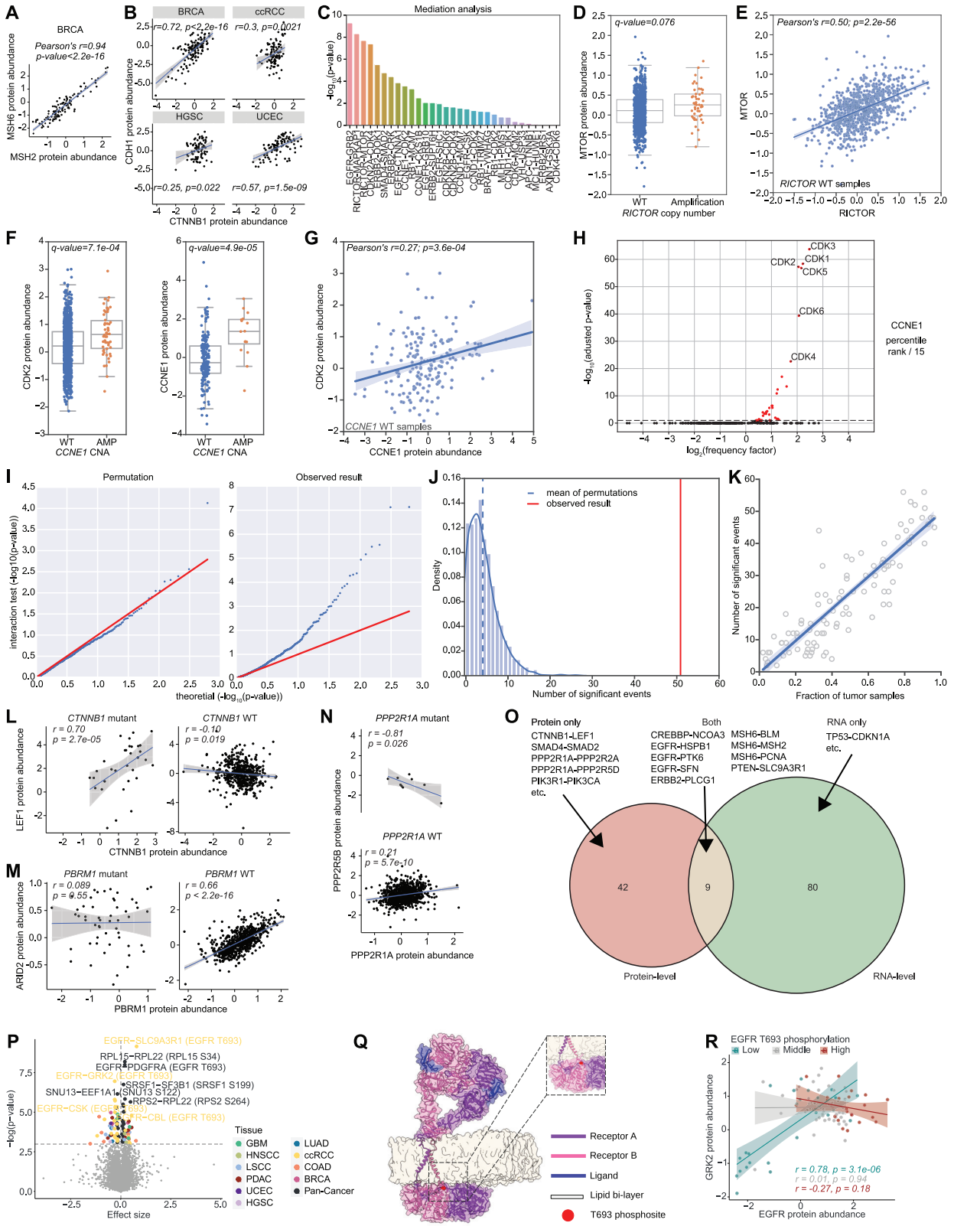
Figure S2. Supporting evidence for *cis*-effect of missense mutations, related to Figure 2

(A) Comparison of ARID1A protein abundance between the matched NATs and WT-tumor samples highlighted the consistency of our original pan-cancer observation at the cohort level for most cancer types. Wilcoxon signed-rank test p values are shown. Boxes represent the interquartile range (IQR, e.g., median, 0.25, and 0.75 quantiles), and whiskers represent the largest and smallest values within the 1.5 × IQR range.

(B) Boxplot showing PTEN phosphatase activity from a saturation mutagenesis screen in Mighell et al.[38] for missense mutations with varying degrees of evidence for oncogenicity in CPTAC. Wilcoxon signed-rank test p values are shown. **, p value < 0.01 and ****, p value < 0.0001. Boxes represent the interquartile range (IQR, e.g., median, 0.25, and 0.75 quantiles), and whiskers represent the largest and smallest values within the 1.5 × IQR range.

(C) Boxplot showing PTEN protein abundance from a saturation mutagenesis screen in Matreyek et al.[39] for missense mutations with varying degrees of evidence for oncogenicity in CPTAC. Wilcoxon signed-rank test p values are shown. *, p value < 0.05. Boxes represent the interquartile range (IQR, e.g., median, 0.25, and 0.75 quantiles), and whiskers represent the largest and smallest values within the 1.5 × IQR range.

(D) Bar plot showing the *cis*-effect of missense mutations that are only computationally predicted to be oncogenic on protein abundance (normalized by RNA expression). The dashed line indicates statistical significance (q value < 0.05).



(legend on next page)

Figure S3. Inference of altered protein-protein interactions through protein co-variation analysis, related to Figure 3

(A) Pearson correlation of protein abundance between MSH2 and MSH6 in breast carcinoma (BRCA). The shaded area indicates the 95% confidence interval of the regression line. The correlation test p value is shown.

(B) Pearson correlation of protein abundance between CTNNB1 and CDH1 in four cancer types. The shaded area indicates the 95% confidence interval of the regression line. Correlation test p values are shown.

(C) Bar plot showing the statistical significance from the mediation analysis for each driver gene (left side of label) impacted by copy-number alterations (CNAs) on the abundance of a protein interactor (right side of label).

(D) Boxplot showing the distribution of mTOR protein abundance stratified by *RICTOR* copy-number status. Wilcoxon signed-rank test q value is shown. Boxes represent the interquartile range (IQR, e.g., median, 0.25, and 0.75 quantiles), and whiskers represent the largest and smallest values within the 1.5 × IQR range.

(E) Pearson correlation between the protein abundance of RICTOR and mTOR in tumor samples without a *RICTOR* amplification.

(F) Distribution of CDK2 (left) and CCNE1 (right) protein abundance stratified by *CCNE1* copy-number status. Wilcoxon signed-rank test q values are shown. Boxes represent the interquartile range (IQR, e.g., median, 0.25, and 0.75 quantiles), and whiskers represent the largest and smallest values within the 1.5 × IQR range.

(G) Pearson correlation between the protein abundance of CDK2 and CCNE1 in tumor samples without a *CCNE1* amplification. Correlation test p values are shown, and r represents the Pearson correlation coefficient.

(H) The Kinase Library enrichment for samples with *CCNE1* amplification vs. WT. Cell-cycle-related CDKs (CDK1-6) are significantly enriched among samples with *CCNE1* amplification. The x axis (log_2FF) indicates the log_2-ratio of the prediction frequency for each kinase between samples with CCNE1 amplification vs. WT, and the y axis indicates the statistical significance based on Fisher's exact test.

(I) Quantile-quantile plots for the p value distribution on permuted (left) or observed (right) data for the interaction term regression model for the influence of driver mutations on protein-protein interactions.

(J) Distribution of the number of significant events from the interaction term regression model on 1,000 permuted data sets (blue) vs. the observed (red line).

(K) Subsampling analysis showing the number of significant events from the interaction term regression model as the data size increases.

(L) Scatterplot showing the correlation between LEF1 and CTNNB1 protein levels, stratified by whether the tumor has an oncogenic mutation in *CTNNB1*.

(M) Scatterplot showing the correlation between ARID2 and PBRM1 protein levels, stratified by whether the tumor has an oncogenic mutation in *PBRM1*.

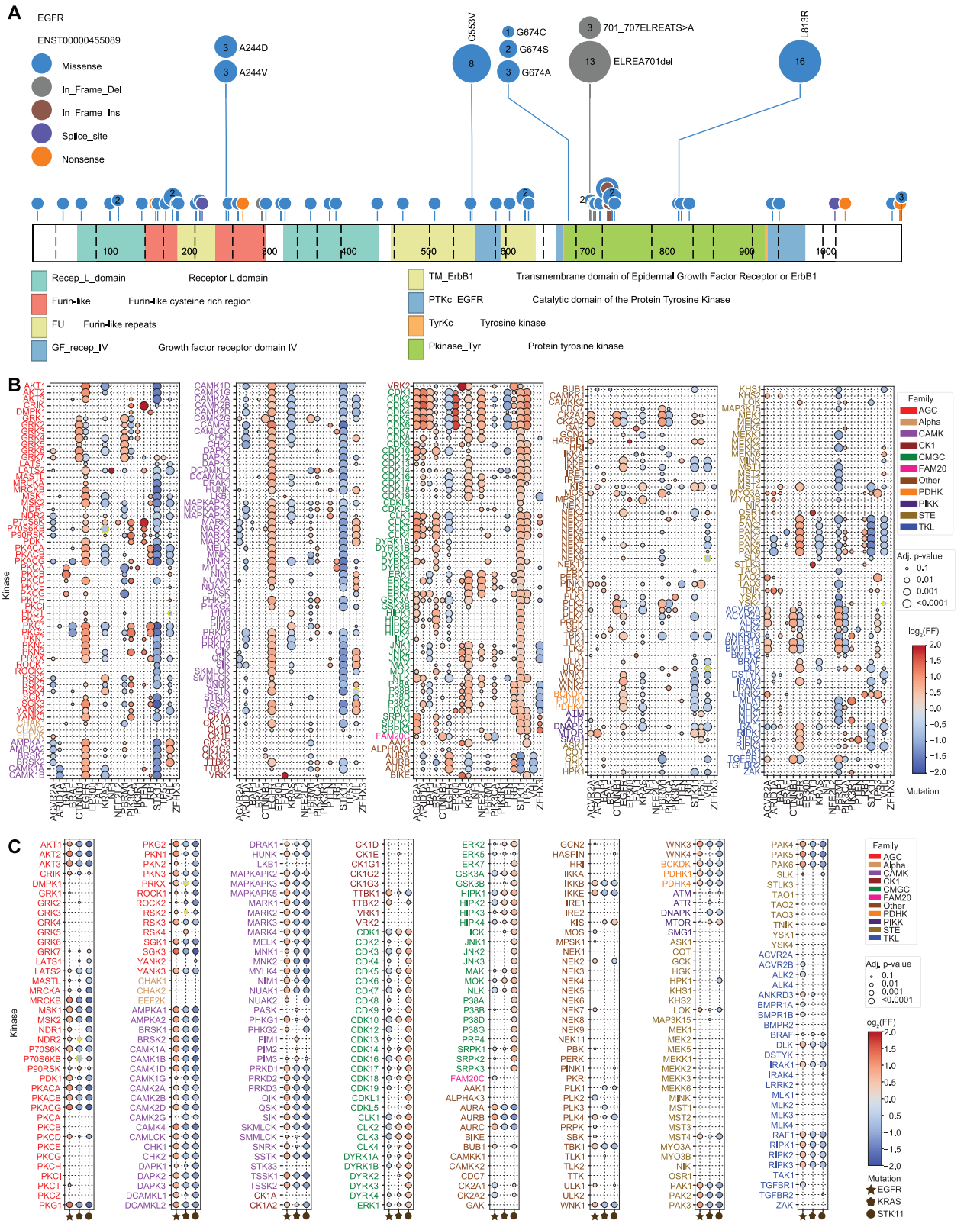
(N) Scatterplot showing the correlation between phosphatase 2A regulatory subunits (PPP2R5D and PPP2R5B) and PPP2R1A protein levels, stratified by whether the tumor has an oncogenic mutation in *PPP2R1A*. Correlation test p values are shown, and r represents the Pearson correlation coefficient in Figures S3L–S3N.

(O) Overlap in significant events from the interaction term regression model when using proteomics vs. RNA-seq data.

(P) Volcano plot to summarize the PPIs significantly influenced by phosphorylation levels at an interface site in one of the corresponding proteins, with the site indicated in parenthesis. Pan-cancer tests accounted for the cohort as a covariate in the model.

(Q) Mapping the EGFR T693 phosphorylation site onto a 3D model of the EGFR homodimer shows that the site falls on the intracellular interface where the dimers interact.

(R) A scatterplot indicates how phosphorylation at EGFR T693 influences the correlation between EGFR and GRK2 protein levels. Correlation test p values are shown, and r represents the Pearson correlation coefficient.



(legend on next page)

Figure S4. Kinase activity analysis of driver genes across different tissues and at the pan-cancer level, related to Figure 4

(A) Lolliplot showing the alterations in *EGFR* observed in CPTAC. Numbers in lollipops indicate the number of tumor samples with that alteration in the entire cohort.

(B) Full Kinase Library enrichment analysis landscape across multiple driver genes at the pan-cancer level. For each kinase and driver, the color of the bubble (log_2FF) indicates the log_2-ratio of the prediction frequency for that kinase between the mutated tumor compared with the WT tumor at the pan-cancer level, and the size of the bubble denotes the statistical significance based on Fisher's exact test.

(C) Full landscape of the Kinase Library enrichment analysis of *EGFR*-, *KRAS*-, and *STK11*-mutated tumors compared to WT in LUAD. For each kinase and driver, the color of the bubble (log_2FF) indicates the log_2-ratio of the prediction frequency for that kinase between the mutated tumor compared with the WT tumor in LUAD, and the size of the bubble denotes the statistical significance based on Fisher's exact test.

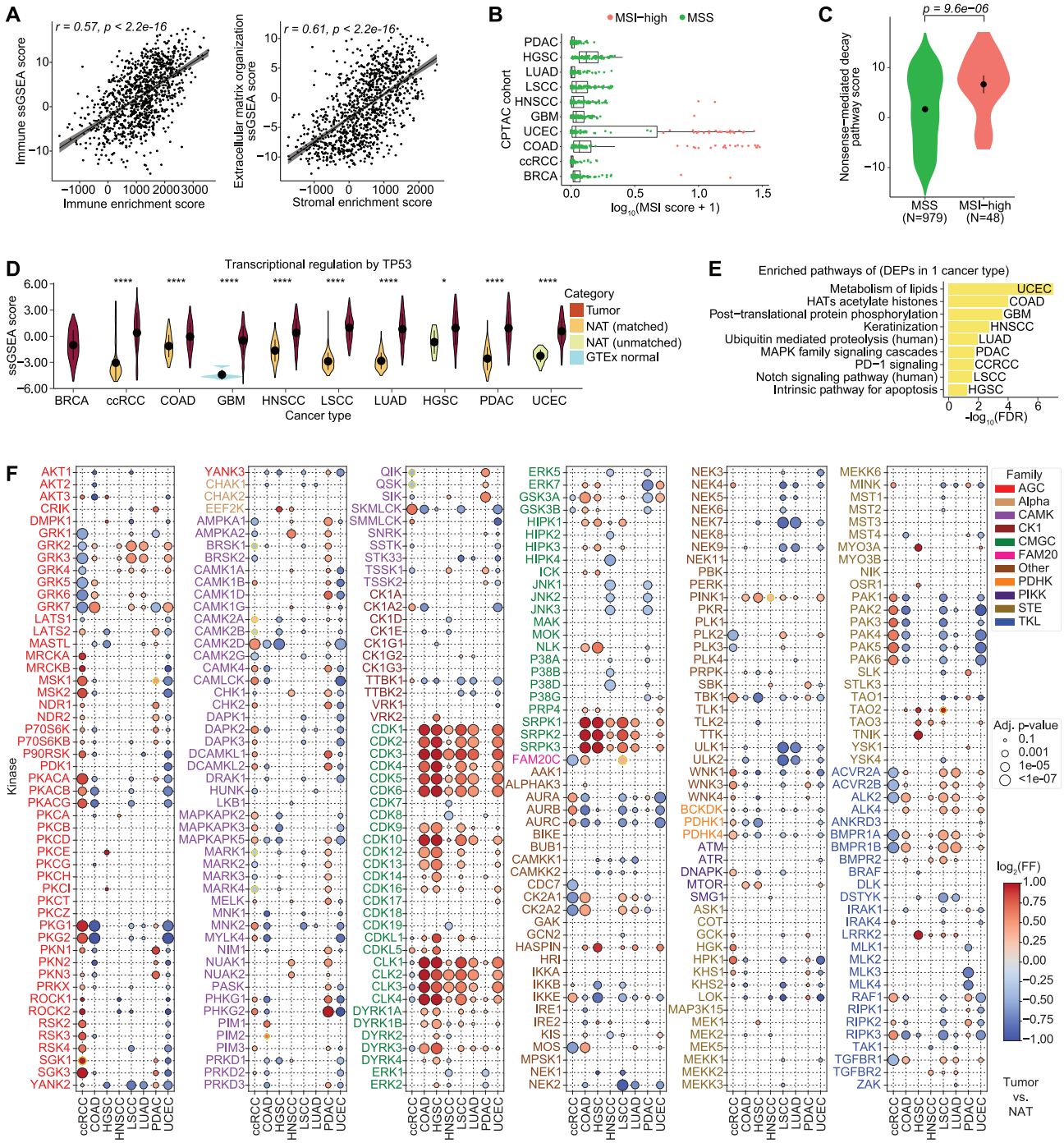


Figure S5. Analysis of normal-adjacent tissue identifies key protein changes for oncogenic pathways, related to Figure 5

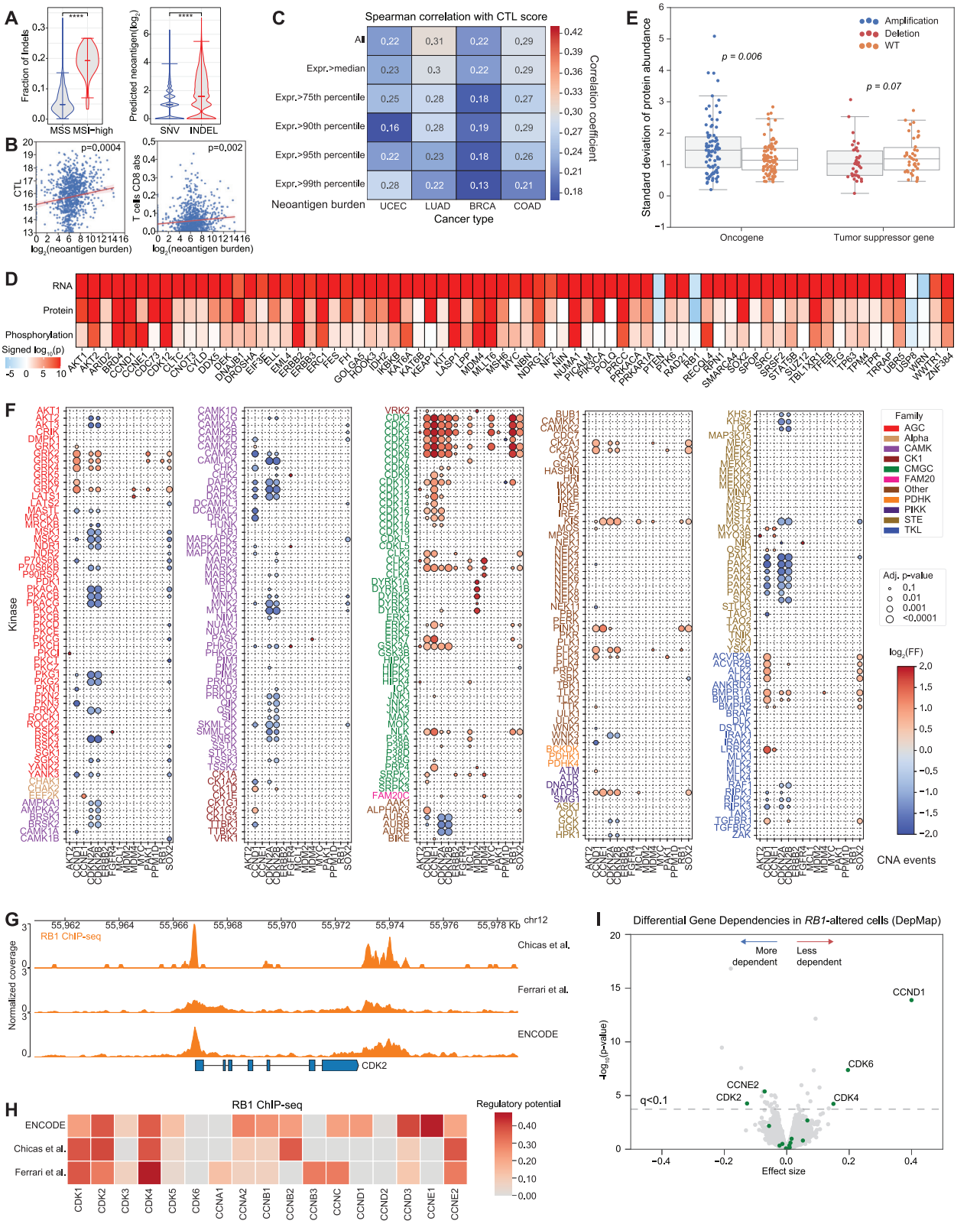
(A) The ssGSEA scores of the immune system and extracellular matrix organization show statistically significant Pearson correlations with the immune score, and stromal score from ESTIMATE, respectively. Correlation test p values are shown, and r represents the Pearson correlation coefficient.

(B) The distribution of the MSI category in each cancer type is displayed in the boxplot colored by the categories of microsatellite stable (MSS) vs. microsatellite instable (MSI-high). Boxes represent the interquartile range (IQR, e.g., median, 0.25, and 0.75 quantiles), and whiskers represent the largest and smallest values within the 1.5 × IQR range.

(C) The nonsense-mediated decay ssGSEA pathway score is enriched in the MSI-high group, followed by the MSS group. Wilcoxon signed-rank test p value of 9.6e−06 is attained.

(legend continued on next page)

(D) Protein-based ssGSEA scores of transcriptional regulation by TP53 are significantly more highly expressed in tumors than that of NATs of all eight cancer types and in GBM compared with GTEx normal samples. Wilcoxon signed-rank test p values are shown. Within the violin plots, the dots represent the mean, and solid lines indicate the error limit defined by the 95% confidence interval. The violin plot outlines demonstrate the kernel probability.
(E) Different enriched pathways in each cancer type based on the DEPs detected only in the corresponding cancer type. Hypergeometric test false discovery rate (FDR)-adjusted p value was attained.
(F) Full landscape of kinase activity across eight cancer types based on the Kinase Library, comparing tumors to their corresponding NATs. For each kinase and tumor type, the color of the bubble (log_2FF) indicates the log_2-ratio of the prediction frequency for that kinase between the tumor and its corresponding NAT, and the size of the bubble denotes the statistical significance based on Fisher's exact test.



(legend on next page)

Figure S6. Analysis of neoantigen burden, kinase activity, and drug sensitivity, related to Figure 6

(A) Left, Violin plot showing the distribution of the fraction of somatic mutations indels for MSS vs. MSI-high tumors. Right, the cumulative number of the predicted neoantigens for single-nucleotide variants (SNV) compared with indels. Wilcoxon signed-rank test p values are shown. ****p value < 0.0001. Within the violin plots, the dots represent the mean, and solid lines indicate the error limit defined by the 95% confidence interval. The violin plot outlines demonstrate the kernel probability.

(B) Left, cytotoxic T lymphocyte (CTL) score correlation with $\log_2$-transformed neoantigen burden. The p value reflects a Wald test after adjusting for cancer type as a covariate. The shaded area represents 95% confidence intervals (left); right, correlation of CD8 T cell score from CIBERSORT (Absolute) with $\log_2$-transformed neoantigen burden. The p value reflects a Wald test after adjusting for cancer type as a covariate. The shaded area represents 95% confidence intervals.

(C) Pearson correlation of neoantigen burden and inferred T cell infiltration (CTL score) for genes expressed at different RNA levels.

(D) *Cis*-effect of copy-number amplifications on the RNA-, protein- and the phosphorylation-level. Red indicates increased expression, while blue indicates decreased expression. Tiles are colored according to the score of the *cis*-effect from red (positive scores) to blue (negative scores). Signed $\log_{10}$(p), multiplying the direction of the coefficient by the $\log_{10}$ of the adjusted p value (capped at ±10).

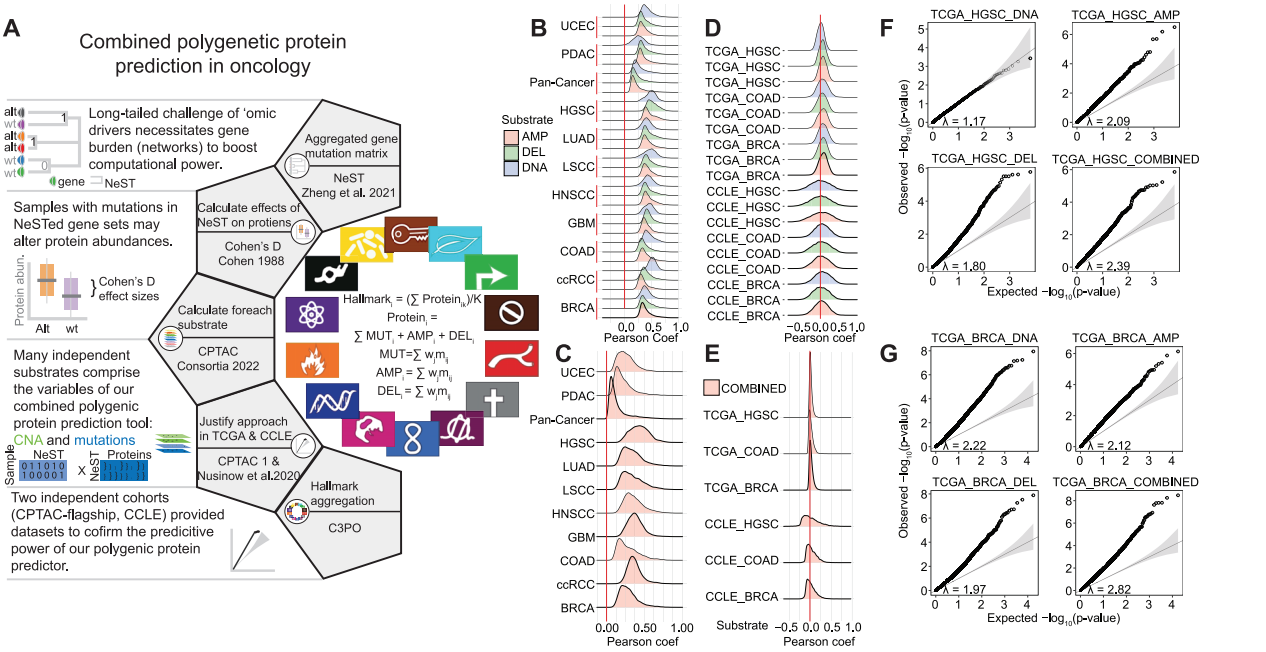
(E) Boxplot comparing the variability in expression for proteins that have amplifications (oncogenes) or deletions (tumor suppressors) compared with wild-type tumors. Each dot represents the standard deviation of protein abundance for a particular oncogene/tumor suppressor. Wilcoxon signed-rank test p values are shown. Boxes represent the interquartile range (IQR, e.g., median, 0.25, and 0.75 quantiles), and whiskers represent the largest and smallest values within the 1.5 × IQR range.

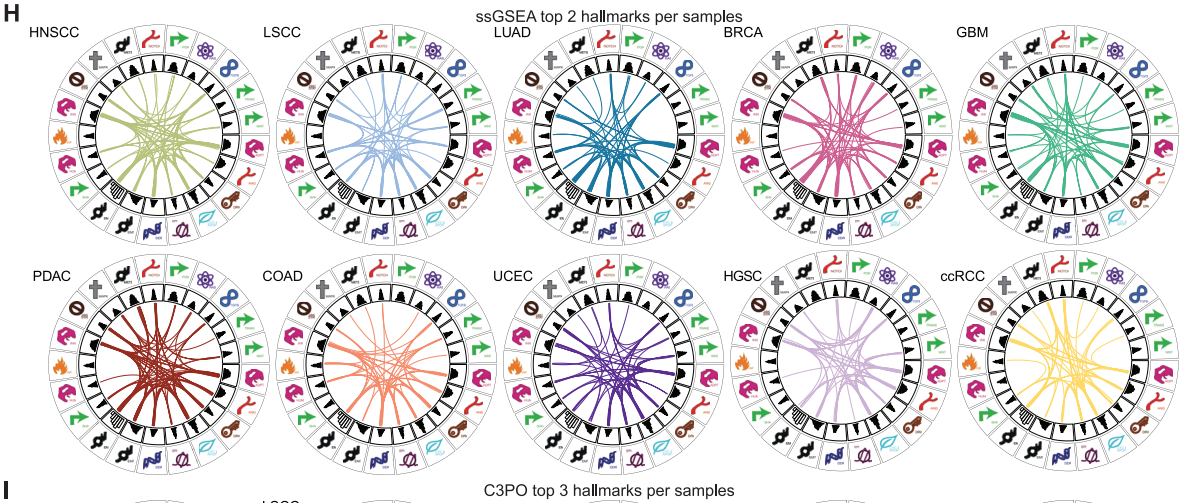
(F) Full landscape of the Kinase Library enrichment results across multiple CNAs at the pan-cancer level. For each kinase and driver, the color of the bubble ($\log_2$FF) indicates the $\log_2$-ratio of the prediction frequency for that kinase between the tumor with the CNA compared to the WT tumor at the pan-cancer level, and the size of the bubble denotes the statistical significance based on Fisher's exact test.

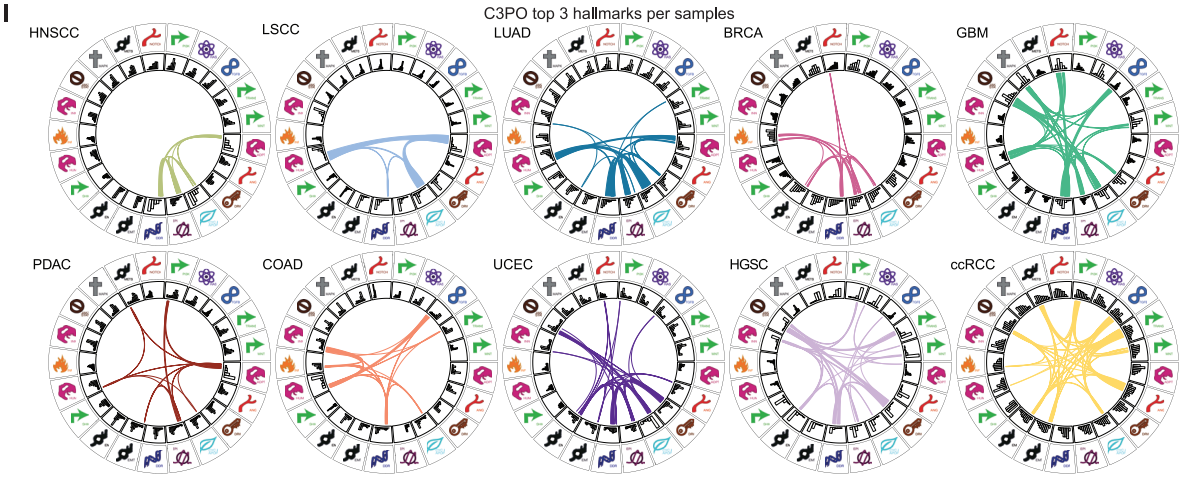
(G) Genome track of the CDK2 locus for RB1 chromatin immunoprecipitation (ChIP) sequencing from three previous studies. Normalized coverage means reads per million.

(H) Heatmap scoring how close RB1 ChIP-seq peaks are to cell-cycle-related genes (CDK1-6, cyclins A–E). The potential regulatory score weights the number and proximity of ChIP-seq peaks to the gene's transcription start site. The potential regulatory score was normalized for each dataset, so the maximum is 1.

(I) Volcano plot showing genes with differential dependencies in *RB1*-altered cell lines from DepMap's CRISPR KO screens. Effect size is the beta coefficient from the linear regression model used in the statistical test. Green dots indicate cell-cycle-related genes (CDK1-6, cyclins A–E).







(legend on next page)

Figure S7. Overview of C3PO and its performance in orthogonal datasets, related to Figure 7

(A) The overview of the C3PO process includes the following step: (1) We used a nested gene relationship architecture to classify altered and unaltered genes. We do 'his to improve upon the long tail of altered genes in cancer; (2) The Cohen's D statistic is used to proxy the effect size in each altered nest and substrate; (3) Effect size matrices are calculated for each genomic substrate (DNA, can amplification, and can deletions); (4) We validated this process in multiple independent cohorts [cancer cell-line encyclopedia (CCLE) and TCGA]; (5) The circle displays how we translate protein activity into hallmark-level scores. The equations within the circle indicate the combined polygenic equations used to generate the various scores.

(B) C3PO protein activity scores were correlated for all proteins measured in CPTAC. Displayed is the distribution of the correlation coefficients attributable to each substrate: DNA (blue), copy-number amplification (AMP, pink), and copy-number deletion (DEL, green) for each cancer type, including the aggregate pan-cancer-trained predictions. The vertical red bar indicates zero correlation.

(C) Combined C3PO protein activity scores were also calculated and correlated with measured protein. Distributions of correlation coefficients are presented for each cancer type, including pan-cancer-trained predictions. The vertical red bar indicates zero correlation.

(D) Similar to (B), C3PO weighted activity matrices were applied to two orthogonal datasets, CPTAC-retrospective samples (TCGA) and 350 cancer cell-line encyclopedia (CCLE) for three cancer types: ovarian (HGSC), breast carcinoma (BRCA), and colorectal tumors (COAD). Similar to (B), predicted protein activity was correlated with measured protein. The vertical red bar indicates zero correlation.

(E) Like (C), combined protein activity predictions were correlated with measured protein. The vertical red bar indicates zero correlation.

(F) QQ plots were used to display all p values generated from Pearson correlations in (D) and (E) for ovarian samples. QQ plots for each substrate were plotted (DNA, top-left; copy-number amplification, top-right; copy-number deletion, bottom-left; and combined scores, bottom-right). Inflated p values were interpreted as the capacity of C3PO to accurately predict protein activity above the expected p value distribution for the number of tests performed. Thus, the higher the p value inflation, the more accurate C3PO predictions are in identifying correlations with genomic alteration and protein activity. Lambda provides the inflations. Lambda (λ) values measure p value distributions above the expected distribution of tests performed.

(G) Similar to (F). QQ plots and lambda values are presented for each substrate in breast cancer.

(H) The circos plots show 21 cancer hallmark pathways for each cancer type. Icons indicate relationships to the original hallmarks outlined by Hanahan and Weinberg[66]. Every sample in CPTAC is represented by the links connecting hallmarks and is colored according to cancer type. Each sample is displayed using one line according to their top two hallmark activity scores generated by ssGSEA.

(I) The circos plots show 21 cancer hallmark pathways for each cancer type. Icons indicate relationships to the original hallmarks outlined by Hanahan and Weinberg[66]. Every sample in CPTAC is represented by the links connecting hallmarks and is colored according to cancer type. Each sample is displayed using three lines according to their top three hallmark activity scores generated by C3PO.